UNIVERSIDAD AUTONOMA
DE MADRID

Escuela Politécnica Superior

Departamento de Ingeniería Informática

PROCESAMIENTO AUTOMATIZADO DE DATOS PROTEÓMICOS: DESDE LA ESPECTROMETRÍA DE MASAS AL CONOCIMIENTO BIOLÓGICO

TOWARDS AN AUTOMATIC PROCESSING OF PROTEOMICS DATA: FROM MASS SPECTROMETRY TO BIOLOGICAL KNOWLEDGE

TESIS DOCTORAL

Juan Alberto Medina Auñón

Madrid, Octubre 2013

UNIVERSIDAD AUTONOMA
DE MADRID

Escuela Politécnica Superior

Departamento de Ingeniería Informática

PROCESAMIENTO AUTOMATIZADO DE DATOS PROTEÓMICOS: DESDE LA ESPECTROMETRÍA DE MASAS AL CONOCIMIENTO BIOLÓGICO

TOWARDS AN AUTOMATIC PROCESSING OF PROTEOMICS DATA: FROM MASS SPECTROMETRY TO BIOLOGICAL KNOWLEDGE

TESIS DOCTORAL

Director	Director	Doctorando
J. P. Albar Ramirez	J. M. Carazo García	J. A. Medina Auñón

Madrid, Octubre de 2013

A C y C

La que siempre me ha acompañado

La que nos acompañará en breve

A mi madre

Por estar siempre

A mi padre

Hoy habrías disfrutado

Agradecimientos

El último capítulo de la tesis siempre se escribe con la satisfacción de haber concluido un apartado de tu vida, pero que sin duda se combina con cierta melancolía cuando te sientas a hacer un repaso de cómo esta etapa ha transcurrido.

Resulta evidente que los agradecimientos tienen que empezar por la familia, ya que esta es la responsable de formarnos como personas, que al fin y al cabo es lo más importante. A Carmen porque es toda mi vida. En todos los años que llevamos juntos siempre nos hemos compenetrado, me ha entregado su amor y sin duda me ha ayudado a ser mejor en todo. Te quiero mi amor. A mi madre, a la que obviamente le debo la vida y todos los esfuerzos que ha demostrado para sacarme adelante y en días como estos es donde un orgullo arraigado en su interior sale a la superficie. ¡Con todo el derecho del mundo, porque para eso es mi madre!. Te quiero mamá. A mi padre, que se que hubiera disfrutado mucho en toda esta etapa y por supuesto en la defensa, pero que sin duda está con nosotros. A mi familia política. ¡Qué buenas verdes con el suegro!. Entre coñas y lances varios siempre nos han apoyado en las distintas aventuras que hemos emprendido. Y por supuesto a mi abuela y mi tía María que en mi interior guardo con celo su recuerdo.

En especial quiero agradecer a los amigos con los que puedo contar. A Nacho por tantas y tantas cosas con sus discusiones y peleas. Cosas normales entre hermanos. A Kike, que le gusta más un café a horas de la siesta que ver perder al Madrid. A los polvorones. ¡Qué Especiales sois!. Con ellos queda una tina pendiente. A Salva y Juanito, que les dejaba encargados de la gerencia del negocio. A Quique, Alicia y Zaira por haber sido nuestra familia de Madrid. A Nacho, que pese a los años y a la distancia siempre que ha podido hemos encontrado un hueco para hidratarnos y contarnos como van las cosas. A Óscar y Samara por tantos y tantos buenos ratos.

Esta etapa de mi vida comenzó hace mucho. Algunos dirían que demasiado. Pero desde entonces y en todo mi recorrido laboral siempre me he sentido respaldado por mis directores. Con ellos, Juan Pablo y José María, me unen vínculos que superan los puramente profesionales y de los cuales me siento orgulloso de haber contado en los distintos retos a los que nos hemos enfrentado. También me gustaría agradecer los prácticos consejos de mi tutor, Roberto, que con una gran efectividad me ha orientado en todos los pasos dados, que no han sido pocos, para llegar hasta este momento. Y a Pablo me gustaría agradecerle que haya sido mi Cicerone por los complicados protocolos de la escuela cuando todo parecía acabado.

No puedo dejar de mencionar mis compañeros del CNB. Alberto y Miguel a los que le he dado la lata con esta memoria y de los cuales he aprendido la experiencia de los que ya han pasado por el trámite. Miguel Ángel y Salva por haber formado un buen equipo de bioinformática que ha permitido colocarnos en el mapa. Los buenos y no tan buenos momentos han dado sus frutos y estoy seguro que seguirán dándolos. A Sergio,

Marisol y Silvia, por ese apoyo mutuo que desde el principio nos hemos brindado y que sin duda transciende de lo puramente proteómico. Al socio, Adán, Carmen y sus flamenquines por esa fuerza y sentido del humor que está presente en la generación más joven del laboratorio. A Fernando por todos esos buenos ratos, algunos sobre temas muy culinarios. A las chicas de ProteoRed, Lola y Virginia, por todas las cosas que han hecho por nosotros y tantos quebraderos de cabeza que nos han resuelto. Y por supuesto a Ana, Rosana, Inés, Gema, Gonzalo, Mari Carmen y Manuel por haber compartido conmigo tantos momentos imborrables. Y en este repaso, tampoco podía pasar por alto la comunidad o familia proteómica de Proteored de la que he aprendido tantas y tantas cosas de este particular mundo. A Fernando, Concha, Félix, Pino, Jesús Mari, Gorka, Vital, Víctor y otros muchos.

Por último, quiero dedicar unas líneas muy especiales a Claudia. Mi niña. Mi vida, cuando esta tesis se esté defiendo a ti te llegarán ciertas ondas que no sabemos muy bien como interpretarás, pero que siempre sepas que esto va de una manera muy especial por ti. En unas semanas te veremos esa carita tan “bonica” que tendrás y te comeremos a besos. Sin duda tu eres el mejor regalo que nos va a traer la vida.

TOWARDS AN AUTOMATIC PROCESSING OF PROTEOMICS DATA: FROM MASS SPECTROMETRY TO BIOLOGICAL KNOWLEDGE

Proteomics covers large-scale studies about proteins, mainly focusing on their structure, functions and interactions. To capture how the proteins are expressed, several high-throughput techniques are available. Protein preparation and separation techniques, mass spectrometry, peptide and protein identification and quantitation are the main stages for a proteomics experiment and they are utilized to obtain a wide picture about how the analyzed proteins are interacting with a particular organism within a specific scenario. The tons of data returned by proteomics experiments require specific bioinformatics approaches to translate these data into information and knowledge.

This thesis is focused on some of these bioinformatics approaches (computational proteomics). Throughout the articles included in this thesis (peer-reviewed), proteomics and computational users will find valuable tools and aids to solve some of the common laboratory lacks. Concretely, the covered areas of interest in the habitual proteomics pipeline are:

- **Proteomics standards data formats**: Two new XML-based data standards are included regarding the electrophoresis separation (GelML) and mass spectrometry quantitation (mzQuantML). Both formats were developed by the HUPO-PSI (HUman Proteome Organization-Proteomics Standards Initiative) working group in accordance with the worldwide standardization proteomics guidelines.

- **Mass Spectrometry validation**: A novel tool (PRIDEViewer) was developed to provide a deep analysis of mass spectrometry results. Data visualization, integration and validation are included to drive users to manage large protein data sets

- **Standard data formats integration and submission**: Specific frameworks (PRIDESpotMapper and ProteoRed MIAPE Web ToolKit, PMWTK) have been developed to include HUPO-PSI data standards into routine laboratory pipelines and to improve proteomics data sharing within the scientific community. Proteomics journals encourage the use of these standards and guidelines to improve the quality of experimental reporting and ease the evaluation and publication of manuscripts.

- **Biological knowledge**: According to the need to solve biological and molecular questions regarding a set of identified proteins, an automatic service (Protein Information and Knowledge Extractor, PIKE) for annotation retrieving has been also included. From a protein data set, this tool extracts and collects the existing biological and functional annotations to get a better understanding of the role played by the proteins in the context of the performed experiment.

In conclusion, the collection of novel works and free bioinformatics utilities enclosed within this PhD project will contribute to improve analytical performance in the majority of proteomics laboratories and open new approaches for both biology and computational research.

Abreviaturas

1D-E	Electroforesis monodimensional o 1D.
2D-E	Electroforesis bidimensional o 2D.
3-DE	Electroforesis tridimensional.
ABRF	Association of Biomolecular Resource Facilities.
ARN	Ácido Ribonucléico.
ARN	Ácido Ribonucléico mensajero.
CAM (Cys)	Carboxyamidomethyl (Cysteine). PTM
CSV	Comma Separated Values.
CV	Controlled Vocabulary.
DAVID	Database for Annotation, Visualization and Integrated Discovery.
DDBJ	Database of Japan.
DIGE	Difference Gel Electrophoresis.
DTT	Dithiothreitol.
EBI	European BIoinformatics Institute.
ECD	Electron Captured Dissociation.
EMBL	European Molecular Biology Laboratory.
ENCODE	The Encyclopedia of DNA Elements.
ESI	Electro-Spray Ionization.
ETD	Electron transfer dissociation.
FMM	File Manager Module.
FTICR	Fourier TRansform Ion Ciclotron Resonance.
FUGE	Functional Genomics Experiment.
Gel-LC-MS	Gel + Liquid Chromatography- Mass Spectrometry.
GO	Gene Ontology.
GOLD	Genome Online Databases.
HPLC	High Performance Liquid Chromatography.
HPLC-MS	High Performance Liquid Chromatography-Mass Spectrometry.
HPP	Human Proteome Project.
HPPP	HUPO Plasma Project.
HPRD	Human Protein Reference Database.
HTML	Hypertext Markup Language.
HUPO	Human Proteome Organisation.
ICAT	Isotope-coded affinity tag.
ICPL	Isotope-coded protein label.
IPI	International Protein Index.
IRM	Information Retrieval Module.
ISCIII-ProteoRed.	Instituto de Salud Carlos III – Instituto Nacional de Proteómica.
iTRAQ	isobaric tag for relative and absolute quantitation.
KEGG	Kyoto Enciclopedia of Genes and Genomes.
LC-MS	Liquid Chromatography- Mass Spectrometry.
Ox (M)	Oxidation (Methionine). PTM
m/z	relación masa – carga.
MALDI	Matrix Assited Laser Desorption/Ionization.
MGF	Mascot General File.
MIAPE	Minimun information About a Proteomics Experiment.

MS	Mass Spectrometry.
MS/MS	Espectrometría de masas tándem.
MudPIT	Multidimensional Protein Identification Technology.
NCBI nr	National Center for Bitechnology Information – non redundant.
NGS	Next Generation Sequencing.
OMIM	OnLine Mendelian inheritance in Man.
OT	OrbiTrap.
PAGE	PolyAcrilamide Gel Electrophoresis.
PCR	Polymerase Chain Reaction.
PDB	Protein Data Bank.
PFF	Peptide Fragment Fingerprinting.
Phos (Ser, Tyr, The)	Phosphorilation (Serine, Tyrosine, Threonine). PTM
pI	Punto isoeléctrico.
PIKE	Protein Information and Knowledge Extractor.
PIR	Protein Information Resource.
PMDB	ProteoRed MIAPE DataBase.
PMF	Peptide Mass Fingerprinting.
PMWTK	ProteoRed MIAPE Web Toolkit.
PRIDE	Proteomics IDEntification database.
PSI	Proteomics Standards Initiative.
PTMs	Post-traductional modifications.
Q	Quadrupole.
Q-TOF	Quadrupole – Time of Flight.
Q-Trap	Quadrople – ion trap.
RP	Reverse-Phase.
SCX	Strong cationic Exchange.
SDS	dodecil-sulfato sódico.
SILAC	Stable Isotope Labelling by Aminoacids in Cell Culture.
SNR	Signal to Noise Ratio.
spHPP	Spanish HPP.
SRM	Selected Reaction Monitoring.
STRING	Search Tool for Retrieval of Interacting Genes/Proteins.
TMT	Isobaric Tandem mass tags.
TOF	Time of Flight.
WFMM	Workflow Manager Module.
XML	eXtensible Markup Language.

ÍNDICE DE CONTENIDOS

BLOQUE 1

BLOQUE 2

BLOQUE 3

Copia artículos compendiados

ÍNDICE DE FIGURAS Y TABLAS

BLOQUE 1: INTRODUCCIÓN

Desde la genómica a la proteómica

1.1. Genoma y Proteoma

En febrero de 2001, dos de las más prestigiosas revistas de investigación, Science y Nature, publican simultáneamente un primer borrador de la secuencia del genoma humano (Lander, Linton et al. 2001; McPherson, Marra et al. 2001; Venter, Adams et al. 2001), lo que constituyó un hito en el camino del entendimiento de la biología. Desde entonces, y gracias a las posibilidades y avances que ofrecen los secuenciadores automáticos de alto rendimiento, se están completando análisis genómicos de multitud de organismos[1], además de refinar y mejorar aquella primera versión del genoma humano (Abecasis, Altshuler et al. 2010). Actualmente y gracias a los equipos de secuenciación de la próxima generación (NGS – *Next generation sequencing*) es factible la secuenciación de un genoma completo en cuestión de días (Wheeler, Srinivasan et al. 2008; Henson, Tischler et al. 2012).

Esta realidad ha permitido disponer de ingentes bancos de datos con la secuencia genética de multitud de organismos. Sin embargo, estas valiosas fuentes presentan ciertas limitaciones, principalmente relacionadas con la falta de información respecto a la función biológica desempeñada por los genes (Claverie 2001; Consortium 2005). Por tanto, aunque no se puede obviar el papel básico que el genoma desempeña

1 El sistema de referencia *Genomes Online Databases* (GOLD: http://genomesonline.org/) cuenta con 6868 organismos secuenciados, 311 de los cuales son eucariotas, 6330 bacterias y 227 arqueobacterias (16 Julio 2013).

para un mejor entendimiento de cualquier organismo y a pesar de iniciativas como el proyecto ENCODE (*The Encyclopedia of DNA Elements*) (Birney, Stamatoyannopoulos et al. 2007) el genoma por sí sólo no proporciona una explicación suficiente de la diversidad los seres vivos y su adaptación al entorno.

El siguiente paso en esta dirección comprende el análisis de la transcripción de los genes en ácido ribonucleico mensajero (ARNm) y cómo estos ARNm dan lugar al conjunto de proteínas de un organismo o proteoma. Este término, acuñado por Mark Wilkins en el congreso de Electroforesis de Siena en 1994 (Wasinger, Cordwell et al. 1995), pretendía establecer un paralelismo semántico a partir del ya descrito genoma, como conjunto de genes que definen a un organismo. Así, el producto último de los genes, las proteínas, definirían el proteoma y al estudio de este último se denomina proteómica.

Sin embargo, el análisis del proteoma es en principio una tarea más laboriosa, compleja y desafiante que el del genoma ya que el objeto de estudio se desplaza a como los genes se expresan en determinadas condiciones ambientales, y especialmente como cambian cuando las condiciones se modifican bien de forma inducida, patológica, fisiológica o lo largo de un proceso de desarrollo.

Este nuevo enfoque implica restricciones técnicas derivadas de condicionantes bioquímicos, sobre todo respecto al amplio rango de concentración en que las proteínas se presentan en distintas partes de la célula o fluidos fisiológicos (p.e 10 órdenes de magnitud en plasma (Anderson and Anderson 2002; States, Omenn et al. 2006), el dinamismo de este rango (Hortin and Sviridov 2010) de unos compartimentos a otros, las modificaciones post-traduccionales (*post-translational modifications*, PTMs) que las proteínas pueden presentar según el papel funcional que desempeñen y las limitaciones intrínsecas asociadas a los instrumentos analíticos tanto en sensibilidad como en rango lineal (ninguna técnica analítica exhibe un rango lineal de detección/cuantificación superior a 5 órdenes de magnitud (Picotti, Bodenmiller et al. 2009)).

Otro condicionante de primera magnitud proviene de la complejidad del proteoma. Para abordar este aspecto las técnicas de separación de proteínas (combinando técnicas electroforéticas y cromatográficas) desempeñan un papel esencial. Además, aunque las células de un mismo organismo comparten el mismo genoma, los proteomas pueden presentar variaciones significativas (Collins 2001). También, podría añadirse la imposibilidad de amplificar el número de copias de las proteínas, al contrario que en los estudios genómicos donde sí es posible generar esta amplificación a través de la reacción en cadena de la polimerasa (PCR – *Polymerase chain reaction*) (Saiki, Gelfand et al. 1988).

Finalmente y como se desarrollará a lo largo de este trabajo, si bien las herramientas computacionales necesarias para la identificación y cuantificación de proteínas así como las bases de datos y repositorios de genes y/o de proteínas han experimentado un desarrollo exponencial en la última década (Benson, Cavanaugh et al. 2013), (Flicek, Ahmed et al. 2013), NCBI_Resource_Coordinators (2013), UniProt_ Consortium (2013), (Vizcaino, Cote et al. 2013), aún precisan de optimización y maduración estadística. Todo ello sitúa el análisis masivo de mezclas complejas de proteínas o proteomas como un reto metodológico y biológico de alta complejidad.

1.2. El objeto del estudio del proteoma: Las proteínas

Las proteínas son macromoléculas poli-funcionales responsables de los procesos esenciales que se producen en un organismo, bien realizando una función individual, como el transporte de moléculas o interviniendo en las reacciones bioquímicas como enzimas, o formando complejos proteicos con una actividad más amplia como por ejemplo la defensa de un organismo.

Respecto a su estructura, una proteína es una secuencia de moléculas simples llamadas aminoácidos, formadas por un átomo central de carbono conocido como Carbono alfa (C_α), al que enlazan, un grupo amino (NH_2), un grupo carboxilo (COOH), un átomo de hidrógeno y una cadena, denominada cadena lateral (R), que caracteriza a cada aminoácido.

En la naturaleza son 20 los aminoácidos más comunes (anexo 1). Estos se concatenan mediante enlaces peptídicos. En cada enlace peptídico formado, el carbono carboxílico del aminoácido a_i se enlaza con el nitrógeno del grupo alfa amino del siguiente aminoácido a_{i+1}. Las sucesivas concatenaciones de aminoácidos forman la denominada secuencia o estructura primaria de la proteína, presentando esta una longitud media de 300 aminoácidos o residuos, aunque la variabilidad es alta, desde menos de 100 a más de 5000.

El enlace peptídico es el responsable de que cada proteína presente una columna vertebral que consiste en la repetición del bloque básico –N–C_α–(CO)–. Consecuentemente, cada proteína presenta un grupo amino en uno de los extremos (N-terminal) de la cadena poli-peptídica y un grupo carboxilo en el otro (C-terminal). En un contexto biológico, la síntesis de una proteína siempre se produce desde el extremo N-terminal al C-terminal. Dicha síntesis, la llevan a cabo los ribosomas que son complejos macromoleculares de proteínas y ácido ribonucleico (ARN) que se encuentran en el citoplasma, en las mitocondrias, asociados al retículo endoplasmático y en los cloroplastos. Los ribosomas sintetizan proteínas a partir de la información genética codificada en el ARN mensajero (ARNm) en un proceso complejo denominado traducción. La figura I1.1 muestra una cadena polipeptídica, donde R_i identifica la cadena lateral y se han omitido los grupos H de los carbonos alfa.

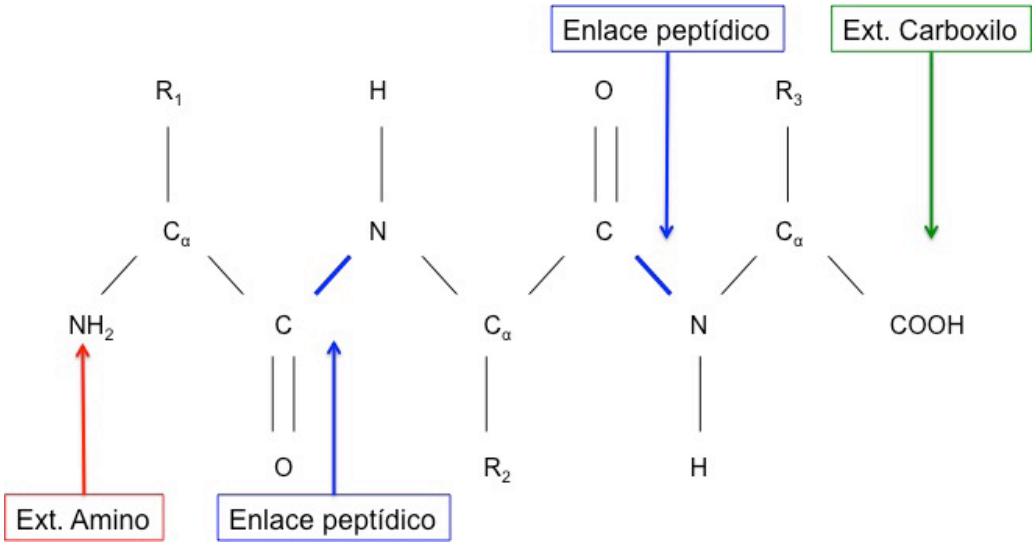

Figura I1.1: Esquema de la estructura primaria de un péptido. Los elementos R_i definen la cadena lateral que a su vez caracteriza al aminoácido. Fuente: (Carlos Setubal 1997).

Finalmente, las proteínas pueden sufrir modificaciones después de su síntesis, conocidas como modificaciones post-traduccionales (PTMs). Las PTMs son uno de los últimos puntos de regulación de la expresión génica y son realmente importantes por la modulación de la función que pueden ejercer sobre la proteína modificada. Aunque existen multitud de tipos (más de 200 modificaciones fisiológicas descritas), algunas de las más frecuentes incluyen acetilaciones (adición de un grupo acetilo), fosforilaciones (adición de un grupo fosfato), metilaciones (adición de un grupo metilo) o glicosilaciones (adición de carbohidratos).

1.3. Estudio del proteoma mediante las técnicas proteómicas

Las técnicas proteómicas permiten el análisis de proteomas complejos. En este sentido, un pilar básico ha sido la espectrometría de masas (MS), una técnica experimental que permite la medición de la masa de iones derivados de moléculas (Price 1991) y facilita los parámetros necesarios para identificar péptidos y proteínas.

1.3.1. Orígenes de la proteómica basada en espectrometría de masas

Durante los años 80 y 90, cambios tecnológicos, fundamentalmente aquellos que permitieron la formación "suave" de iones sin rotura descontrolada de su estructura, permitieron la aplicación de los espectrómetros de masas al análisis de macromoléculas biológicas (Biemann 1992). Dichos cambios fundamentales fueron: el desarrollo de métodos de ionización por electro-nebulización o electro-spray (ESI[2]) (Fenn, Mann et al. 1989) y la ionización/desorpción inducida por láser y asistida por matriz[3] (MALDI[4]), (Karas and Hillenkamp 1988).

1.3.2. Aproximaciones analíticas basadas en espectrometría de masas

Conceptualmente, las aproximaciones que modelan el flujo de trabajo en proteómica basado en espectrometría de masas son: Arriba-Abajo o *top-down* y Abajo-Arriba o *bottom-up* (Fig. I1.2). En la primera o *top-down*, proteínas intactas o fragmentos relativamente grandes son ionizados y analizados. El análisis de la "estructura/composición" se lleva a cabo mediante métodos especializados de fragmentación de secuencias poli-peptídicas grandes como la disociación por captura de electrones (*electron capture dissociation, ECD*) ((Zubarev, Horn et al. 2000)) o la disociación por transferencia de electrones (*electron transfer dissociation, ETD*) ((Syka, Coon et al. 2004) permitiendo la identificación de proteínas bajo determinadas condiciones.

---

2 ESI: ElectroSpray Ionization.

3 Material orgánico de bajo peso molecular que se utiliza para proteger a la biomolécula de ser destruida y para facilitar su vaporización e ionización.

4 MALDI: Matrix-Assisted Laser Desorption/Ionization.

Por su parte, en la estrategia *bottom-up* las proteínas son previamente fragmentadas generando una mezcla compleja de péptidos. Este proceso se realiza por digestión proteolítica mediante una enzima[5] con actividad endoproteasa y especificidad de rotura conocida, que convierte la proteína en péptidos con residuos de aminoácidos conocidos en alguno de sus extremos.

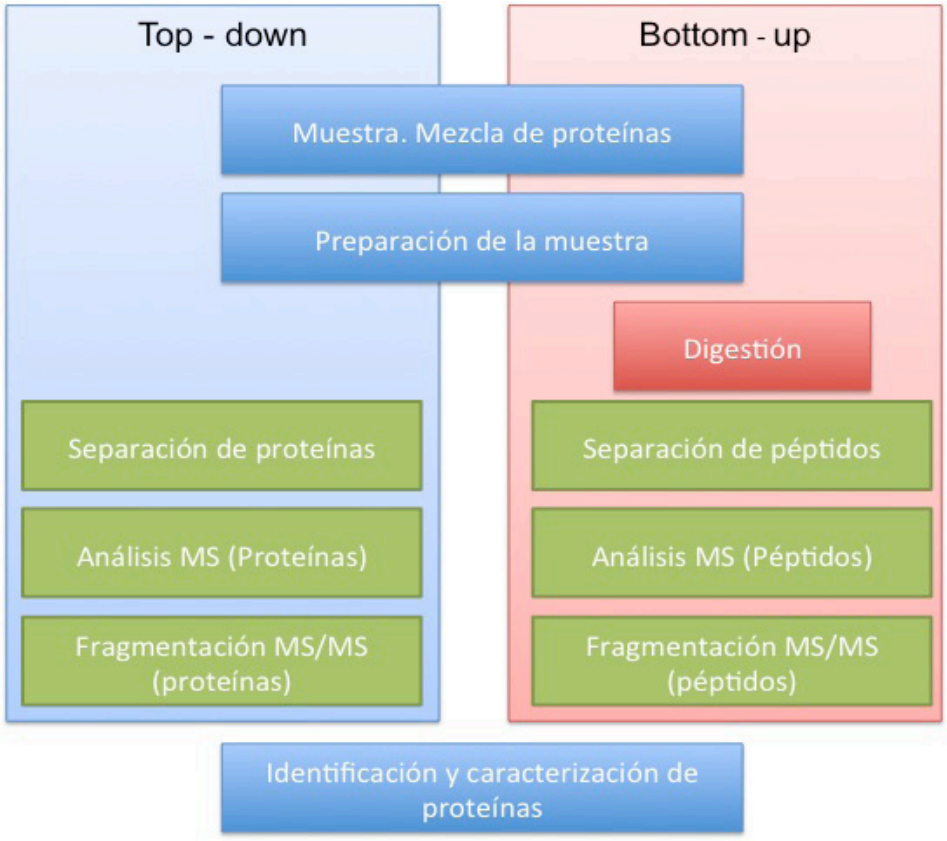

Figura I1.2:

Representación de las dos aproximaciones utilizadas en análisis proteómicos.

Izquierda: Análisis de proteínas intactas o fragmentos relativamente grandes (top-down).

Derecha: Análisis de proteínas a partir de sus péptidos (bottom-up).

Adaptación de la figura 2.2 (pag. 20) de Foodomics, Alejandro Cifuentes, 2013. Willey.

Desde un punto de vista analítico-bioquímico sería ideal analizar proteínas completas, sin embargo, limitaciones en la resolución y precisión de medida por parte de los espectrómetros de masas y de las técnicas de fraccionamiento a nivel de proteína, hacen que a día de hoy la estrategia de análisis de *bottom-up* sea la más usada.

En este último caso, a pesar del incremento del número de elementos a analizar, y por tanto de la complejidad, la instrumentación analítica está más optimizada para biomoléculas de relativamente bajo peso molecular tanto para su separación (cromatografía de alta resolución) como para su identificación (espectrometría de masas), permitiendo un análisis más optimizado que con proteínas intactas. Sin embargo, también presenta ciertas limitaciones en términos de cobertura de secuencia de las proteínas identificadas, detección de PTMs y ambigüedad en la asignación de péptidos redundantes (Yates, Ruse et al. 2009).

En las siguientes secciones se detallarán las distintas fases en las que se puede dividir la estrategia *bottom-up*. El flujo de trabajo que se establece para el denominado análisis masivo o no dirigido (*MS-based shotgun proteomics o unbiassed proteomics*) de mezclas complejas de proteínas comprende: 1) extracción y separación de proteínas, 2) adquisición e identificación por espectrometría de masas y 3) análisis de los datos y generación de resultados.

5 Dependiendo de la enzima, la proteólisis genera péptidos distintos. En proteómica, La más utilizada es la tripsina, y en la mayoría de los casos hidroliza los enlaces peptídicos tras el carbono carboxílico de residuos de lisina (K) o arginina (R). Los péptidos sometidos a digestión con tripsina se denominan péptidos trípticos y contienen por tanto K o R en su extremo carboxilo, con la única posible excepción del extremo C terminal de la proteína.

1.4. Técnicas de separación de proteínas

La gran complejidad de las muestras biológicas obliga a emplear potentes técnicas de extracción y separación previas al análisis por espectrometría de masas. Estas técnicas analíticas sirven en primer término para adecuar el material de partida al análisis al que va a ser sometido en el espectrómetro de masas, pudiéndose procesar fluidos, líneas celulares o tejidos, requiriendo cada tipo un protocolo propio de extracción. En segundo término, hay que destacar que una muestra no se compone de proteínas aisladas sino de una mezcla compleja de estas y otras biomoléculas y analitos.

Una vez procesada la muestra (preparación de la muestra), ésta debe someterse a un proceso de separación o fraccionamiento de sus componentes para, en la medida de lo posible, realizar análisis individualizados de cada molécula de interés, péptido o proteína, o para un conjunto significativamente reducido de éstas.

1.4.1. Electroforesis mono y bidimensional

Estas técnicas separan las proteínas presentes en una muestra según las propiedades físico-químicas de carga y masa molecular. La carga neta de una proteína varía en función de su punto isoeléctrico (pI) y el pH de la solución que contiene la proteína. El punto isoeléctrico se define como el pH para el que la carga neta de la proteína es 0. Es decir, cuando el pH del medio es igual al pI la proteína es eléctricamente neutra. Por su parte, la masa molecular de una proteína es la sumatoria de las masas individuales de los residuos de aminoácido que la forman. Por tanto, ambas propiedades vienen determinadas por la secuencia de aminoácidos y las PTMs de cada una de las proteínas presentes en la muestra.

La separación por electroforesis se categoriza si sólo 1 o las 2 de las dos propiedades citadas anteriormente se consideran (Lieber 2001). La primera de estas técnicas, electroforesis mono-dimensional en gel de poliacrilamida en presencia de dodecil-sulfato sódico (SDS-PAGE) (Laemmli 1970), se basa en la separación por masa molecular. En esta técnica, el extracto proteico se disuelve en un tampón de aplicación que usualmente contiene un agente reductor como DTT (Dithiothreitol) (Cleland 1964) para reducir los puentes disulfuro y el detergente aniónico SDS. Este compuesto se une a la proteína en proporción constante (aproximadamente una molécula de SDS por cada dos aminoácidos) y permite normalizar el comportamiento electroforético de las proteínas con masa molecular similar. El SDS actúa rompiendo enlaces no covalentes en las proteínas, desnaturalizándolas y provocando que pierdan su conformación nativa. Este complejo “SDS-proteína” presenta una estructura nueva y común a todas las proteínas en forma de varilla (*rod-like*) siendo la longitud de esta varilla la que determina el grado de fricción a través de los poros que constituyen el entramado del gel de poliacrilamida y por tanto su desplazamiento en el campo eléctrico. El ratio masa/carga es el mismo para todos los complejos SDS-proteína, siendo el coeficiente de fricción el que determina su movilidad en este tipo de electroforesis. Cuando son sometidos a un campo eléctrico los complejos SDS-proteína migran a través de poros existentes en el gel a diferente velocidad en función de su tamaño (Fig. I1.3). La tinción posterior del gel, mediante de compuestos colorantes, como el azul de Coomassie (*Coomasie blue*) (Fazekas de St Groth, Webster et al. 1963), permite visualizar las proteínas en forma de bandas dentro del gel. La masa molecular de una proteína puede determinarse por comparación de su movilidad electroforética con la de un panel de proteínas de masa molecular conocida, llamadas patrones.

La segunda técnica de separación tradicionalmente empleada en proteómica es la electroforesis bidimensional (2D-E). Esta técnica tiene un poder de resolución mucho mayor que la 1D-E al combinar dos procesos de separación, uno basado en la carga neta de las proteínas y el otro basado en su tamaño. Esta técnica es utilizada cuando se requiere obtener una buena resolución en muestras proteicas complejas.

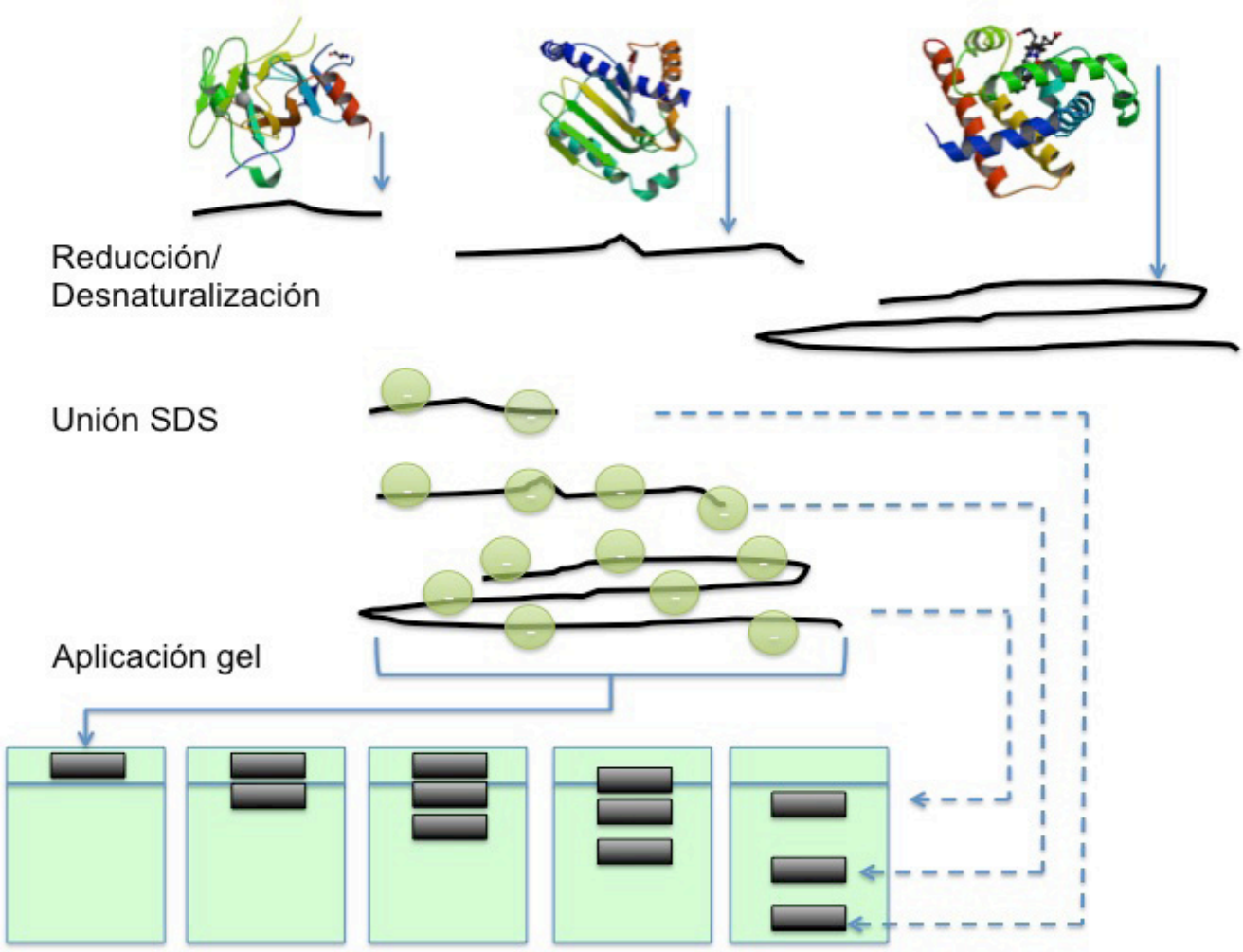

Figura I1.3: Representación esquemática 1D-SDS-PAGE. Adaptación de (Lieber 2001). Fig. 2. Pág 35.

En 2D-E (O'Farrell 1975), el extracto proteico es primero separado mediante isoelectroenfoque (IEF) (Horth, Miller et al. 2006) en un gradiente de pH generado por la presencia de anfolitos, generalmente anclados a un soporte sólido como un gel de poliacrilamida (*gel strip*) permitiendo la separación de las proteínas según su punto isoeléctrico. Para permitir la migración de proteínas en este gradiente es necesario someter al extracto proteico a un campo eléctrico. Debido a que la tira de gel se dispone de modo que la región próxima al ánodo (polo positivo) es más ácida que la que contacta con el cátodo (polo negativo), aquellas proteínas que se encuentran en regiones de pH inferior a su punto isoeléctrico estarán cargadas positivamente y migraran hacia el cátodo, mientras que las que se encuentran en zonas con pH más altos que su punto isoeléctrico tendrán carga negativa y migrarán hacia el ánodo.

La migración las conducirá a una región donde el pH coincidirá con su punto isoeléctrico, resultando una carga neta nula y por tanto eliminando su movilidad. De esta forma las proteínas se sitúan en estrechas bandas donde coincide su punto isoeléctrico con el pH. Una vez separadas las proteínas por focalización

isoeléctrica, las tiras de gel son transferidas a un medio que contiene SDS y sometidas a una segunda separación en un gel de poliacrilamida-SDS según su masa molecular, en un procedimiento similar al descrito para 1D-SDS-PAGE (Fig. I1.4).

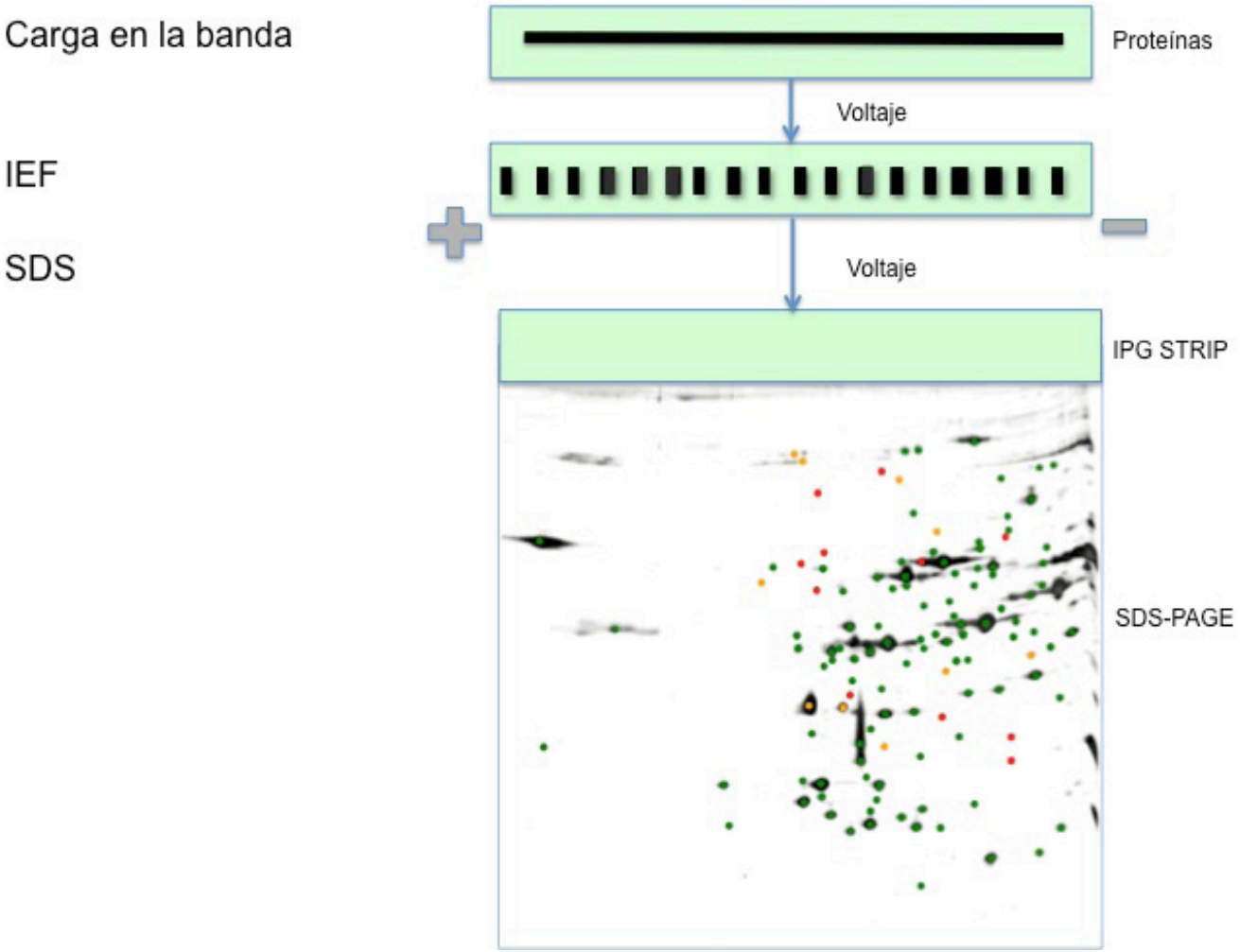

Figura I1.4: Representación esquemática de 2D-SDS-PAGE. Adaptación de (Lieber 2001). Fig 3. Pag 37.

Por último y debido a que las proteínas contenidas en el gel no son visibles a simple vista, se requiere el paso adicional de la tinción donde se revelará la presencia de las proteínas tal y como se observa en la figura. Las manchas más oscuras que el fondo reciben el nombre de *spots*. Un gel simple de estas características en un rango amplio de pH (3-10) permite visualizar varios miles de spots. Entre los agentes más utilizados encontramos el colorante azul de Coomasie o el nitrato de plata, que en condiciones especiales permite el posterior análisis por espectrometría de masas (Shevchenko, Wilm et al. 1996). Otra opción para poder visualizar las proteínas separadas es el marcaje o tinción con algún fluoróforo o reactivo fluorescente que permita detectar los spots mediante la emisión de fluorescencia. Los ejemplos más típicos son el reactivo Sypro (Yan, Harry et al. 2000) en sus distintas variantes y los reactivos basados en cianinas (Cy2, Cy3 y Cy5) característicos del marcaje DIGE (Unlu, Morgan et al. 1997).

Habitualmente el gel es escaneado o fotografiado para obtener una imagen digital del mismo que permita su análisis informático con el fin de determinar las coordenadas, volumen e intensidad de los spots de

interés. Siguiendo con el flujo clásico proteómico los spots seleccionados serán posteriormente recortados, las proteínas que contienen digeridas en el gel con endoproteasas específicas, los péptidos generados extraídos y finalmente analizados mediante espectrometría de masas.

1.4.2. Cromatografía líquida de alta resolución (HPLC) (*High Performance Liquid Chromatography*)

La cromatografía es una técnica analítica y/o preparativa que separa una mezcla de solutos basándose en la diferente velocidad de desplazamiento que estos experimentan al ser arrastrados por una fase móvil líquida a través de un lecho cromatográfico que contiene una fase estacionaria sólida (Miguel Valcárcel Cases 1988).

La separación se basa en diferentes tipos de interacciones entre las sustancias analizadas y el lecho cromatográfico en función de alguna propiedad físico-química como puede ser el tamaño, la carga, la polaridad o la hidrofobicidad de los componentes. En el caso de proteómica, se puede utilizar tanto para la separación de proteínas como de péptidos, siendo esta última la usada con más frecuencia. Las diferentes propiedades utilizadas como base para la separación (polaridad, hidrofobicidad o carga) se utilizan en combinación con distintas configuraciones de fases estacionaria y móvil derivando en diferentes tipos de cromatografía que puede emplearse para la separación de péptidos y proteínas (tabla I1.1).

Tipo	Propiedad	F. Estacionaria	F. Móvil	Comentarios
Fase reversa	Polaridad	Apolar	Polar moderada	Muy utilizada
Intercambio iónico	Carga	Grupos funcionales con carga opuesta al analito.	Aumento de la fuerza iónica o cambio de pH	Muy utilizada como primera dimensión
Afinidad	Interacción específica con un ligando	Ligando de alta afinidad con el analito.	Aumento de fuerza iónica, cambio de pH, cambios de polaridad, agéntes caotrópicos o ligandos competitivos.	Muy utilizada
Fase normal	Polaridad	Polar	Apolar	En desuso

Tabla I1.1: Tipos de HPLC más empleados en proteómica.

En estudios proteómicos se suelen combinar varias separaciones ortogonales donde el último paso es una cromatografía líquida de fase reversa (Dong 1992). El acoplamiento de este tipo de cromatografía con un espectrómetro de masas de recibe el nombre de LC-MS o HPLC-MS (Washburn, Wolters et al. 2001). La integración de un equipo de cromatografía líquida de fase reversa con espectrómetros de masas en tándem MS/MS proporcionan una plataforma con demostrado rendimiento para la separación e identificación de péptidos (Holcapek, Jirasko et al. 2012). En esta configuración, los péptidos eluídos del cromatógrafo son casi instantáneamente ionizados (normalmente por electronebulización), introducidos, aislados y analizados en un espectrómetro de masas. El resultado se representa mediante un cromatograma en el que los componentes, llamados picos, representan los péptidos que han sido aislados y posteriormente analizados por el espectrómetro de masas.

1.5. Técnicas de Identificación de proteínas por espectrometría de masas

Una vez que la muestra de partida se ha sometido al proceso de separación, los péptidos o proteínas se detectan e identifican mediante espectrometría de masas. Existen diversas aproximaciones experimentales, entre las cuales la más frecuente es el análisis conocido como *Shotgun proteomics*, que permite el análisis de muestras extremadamente complejas (proteomas o subproteomas) mediante la combinación de técnicas de separación con distintas configuraciones de espectrómetros de masas. Se emplean dos configuraciones instrumentales típicas que combinan separación bidimensional seguida de análisis por espectrometría de masas en tándem o MS/MS. La primera configuración, conocida como MudPIT (*Multidimensional Protein Identification Technology*), parte de un digerido proteico (conjunto de péptidos), que es fraccionado en una primera separación cromatográfica por intercambio catiónico fuerte (SCX) seguida por una separación nano-cromatrográfica por fase reversa (RP-nanoLC), acoplada esta última al espectrómetro de masas (Zhang, Fang et al. 2010). La segunda configuración empleada en análisis masivos es la conocida como Gel-LC-MS, donde las proteínas son primero separadas en bandas del gel de poliacrilamida-SDS, recortadas, digeridas en gel y posteriormente los péptidos obtenidos separados en un nano-cromatógrafo líquido que está acoplado a un espectrómetro de masas.

La principal ventaja del análisis masivo es su robustez y reproducibilidad. Además el conjunto de proteínas detectadas son identificadas en base a múltiples péptidos, lo que le confiere una mayor confianza. El mayor inconveniente de esta aproximación es que incrementa la complejidad del análisis debido fundamentalmente al gran número de péptidos que tienen que ser analizados y también, como se describirá en próximas secciones, la pérdida de referencia directa con la proteína intacta.

1.5.1. El espectrómetro de masas. Configuración

El principio básico de la espectrometría de masas es la generación de iones que son separados en base a su relación masa-carga (m/z) y posteriormente detectados. A pesar de la gran heterogeneidad existente, los espectrómetros de masas se componen de 3 elementos fundamentales: Fuente de ionización, analizador y detector (Fig. I1.5).

El primero de los componentes, la fuente de ionización, permite conferir una carga eléctrica o ionizar al analito y llevarlo a fase gaseosa, haciéndolo así susceptible a ser medido. Las técnicas dominantes en proteómica, han sido los métodos de ionización por electro-nebulización o electro-spray (ESI) (Fenn, Mann et al. 1989) y la ionización/desorpción inducida por láser y asistida por matriz (MALDI), (Karas and Hillenkamp 1988). La ionización puede tener lugar en modo positivo (generación de cationes, iónes con carga positiva) o negativo (generación de aniones, cargados negativamente). En proteómica, en la inmensa mayoría de los casos se trabaja en modo positivo de modo que los péptidos o proteínas ganan uno o más protones para convertirse en sus respectivos cationes. Una vez que el analito ha sido ionizado pasa al analizador. Este componente establece las condiciones necesarias mediante campos eléctricos y/o magnéticos para separar los analítos en función de su m/z.

El analizador se puede utilizar tanto para seleccionar una ventana reducida de iones de m/z similar como para explorar una colección de estos y permitir su catalogación en base a su m/z. En cuestión de tipos, los más comunes son el tubo de vuelo (*Time of Flight:* TOF), la trampa de iones (*Ion-Trap*), el cuadrupolo (quadrupole), el triple cuadrupolo (QQQ), las basadas en la transformada de Fourier Orbitrap® (OT) e Ion Ciclotron (FTICR) o bien configuraciones de varios analizadores híbridas (cuadrupolo-tiempo de vuelo, Q-TOF, o cuadrupolo-trampa de iones, Q-Trap).

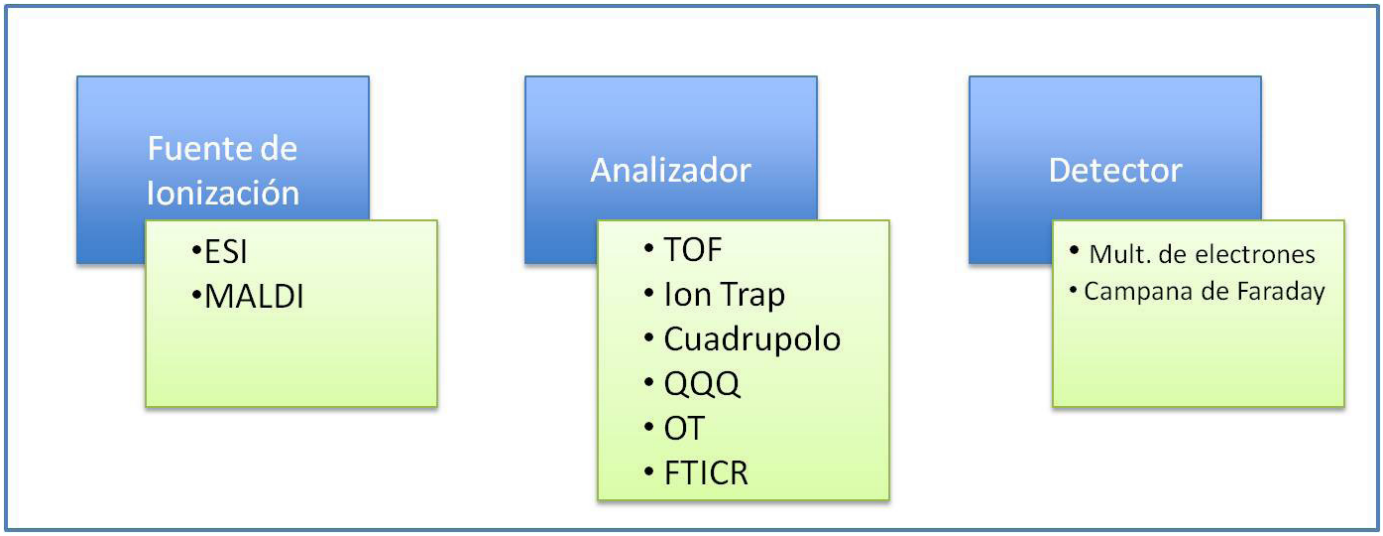

Figura I1.5: Configuración del espectrómetro de masas y principales componentes utilizados en proteómica.

Por último el detector registra la carga inducida o la corriente producida cuando un ion pasa cerca o golpea una superficie. La señal que obtiene este elemento final del espectrómetro de masas es registrada y traducida en el denominado espectro de masas. Los multiplicadores de electrones o fotomultiplicadores son los detectores más utilizados, aunque se han empleado otros detectores como las tazas de Faraday. En el caso de los analizadores basados en la transformada de Fourier, el detector se integra en el analizador y consiste en un par de láminas metálicas que registran una débil corriente que se genera en un circuito colocado entre las placas cuando el ion pasa por éstas.

1.5.2. Espectrometría de Masas en Tándem (MS/MS). Fragmentación de péptidos

En la espectrometría de masas en tándem o MS/MS dos analizadores se combinan para en un primer término aislar un péptido y posteriormente, en el segundo analizador, ser fragmentado. El conjunto de iones correspondientes a los fragmentos obtenidos se representa en un espectro de fragmentación o espectro MS/MS y permite determinar la secuencia de aminoácidos del péptido (Aebersold and Goodlett 2001). Simplificando el proceso a la fragmentación de un único péptido, ésta se lleva a cabo primordialmente en los enlaces peptídicos, generando las denominadas series de fragmentación (Fig I1.6). Estas se nombran (Roepstorff and Fohlman 1984) atendiendo a dos elementos: 1) la posición dentro de la cadena peptídica donde se produce la fragmentación y 2) el extremo del péptido que conserve el fragmento.

En el caso de conservar el extremo amino, las series generadas son a, b o c, mientras que el caso contrario, donde permanece el extremo carboxilo, las series generadas son x, y o z. De estas series las que se generan por la ruptura del enlace peptídico son la serie b en el caso de conservar el extremo amino, y la serie y si cuenta con el extremo carboxilo. En el caso ideal cuando se produce la fragmentación, se generan dos fragmentos complementarios cada uno de los cuales con su extremo correspondiente intacto. Si se fragmentase un péptido de longitud l, los fragmentos complementarios obtenidos en un determinado punto n de la cadena peptídica, serían b_n e y_{l-n} (p.e en la figura b_1-y_2 o b_2-y_1). Para que los fragmentos originados sean detectados, estos tienen que estar cargados, es decir, formar iones. Este estado de carga puede ser múltiple, pudiendo, un ion fragmento presentar carga 1+, 2+, 3+, 4+, siendo esta igual o menor a la carga del péptido precursor que se está fragmentando dependiendo del número de protones retirados (carga negativa) o transferidos (carga positiva).

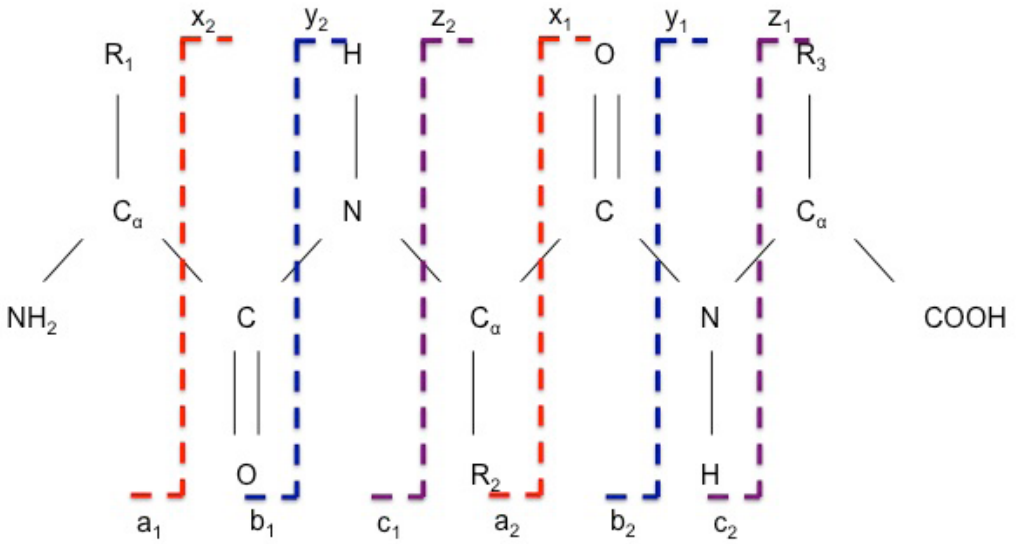

Figura 1.6: Esquema de fragmentación de un péptido. El proceso genera una serie de iones que componen las series de fragmentación y que siguen la nomenclatura propuesta por (Roepstorff and Fohlman 1984). En la fragmentación, se generan dos fragmentos complementarios, con un estado de carga, y mediante los cuales se permite determinar la secuencia de aminoácidos del péptido fragmentado. Fuente: MatrixScience. www.matrixscience.com

Para determinar la secuencia de aminoácidos a partir de los iones generados durante la fragmentación es necesario identificar aquellos iones que pertenezcan a la misma serie de fragmentación (por ejemplo: serie y) y calcular las diferencias de masa, atendiendo su relación m/z, entre los iones. La diferencia entre las relaciones m/z de dos iones consecutivos de la misma serie de fragmentación coincide con el cociente entre la masa neutra del residuo de aminoácido ubicado en dicha posición dividida por el estado de carga de iones que delimitan dicho residuo. Esta es la base de los sistemas de secuenciación *de Novo* (Roepstorff and Fohlman 1984; Biemann and Martin 1987) para la identificación de péptidos que no han sido registrados en una base de datos y también para la localización concreta de residuos modificados por alguna PTM.

Además de las series fragmentación pueden aparecer otros fragmentos (Biemann 1990) correspondientes a pérdidas de agua o amonio, iones imonio, fragmentos internos o iones satélite (Biemann and Martin 1987; Ambihapathy, Yalcin et al. 1997), complicando el espectro de fragmentación o MS/MS (fig I1.7), aunque frecuentemente la información que aportan es útil en la identificación.

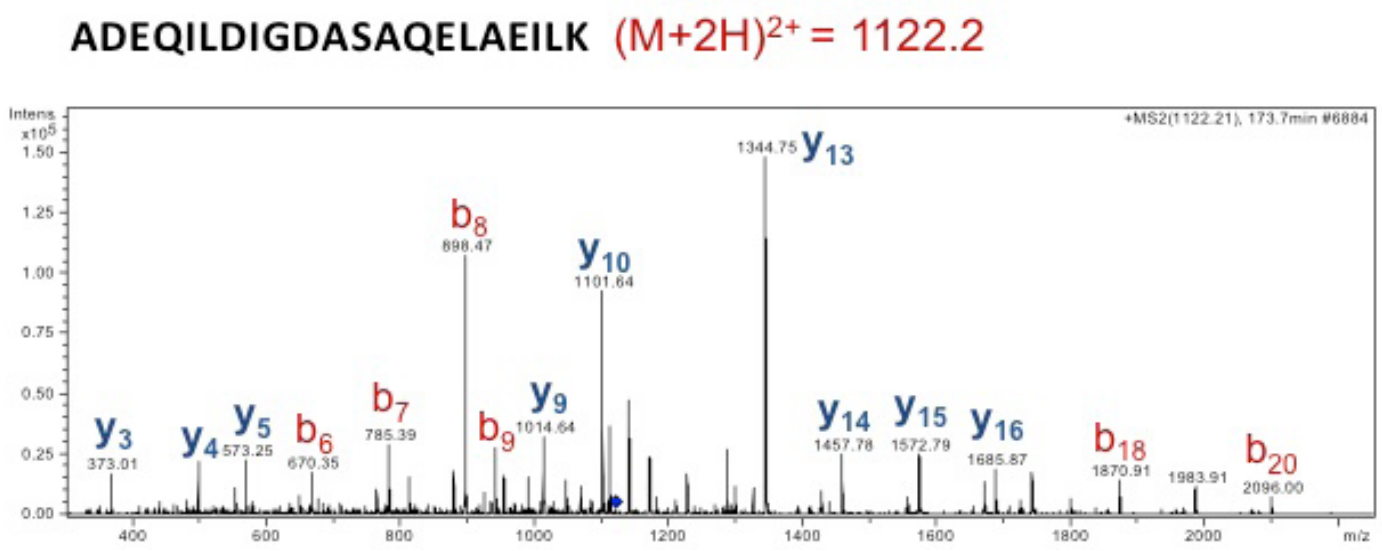

Figura I1.7: Espectro de fragmentación o tándem o MS/MS. En este caso los picos corresponden a iones generados durante el proceso de fragmentación y atendiendo a las posiciones posibles de ruptura del péptido. También pueden aparecer iones relativos a pérdidas, fragmentos internos, iones imonio o iones satélite. *Fuente:* Espectro obtenido en el departamento de proteómica del Centro Nacional de Biotecnología – CSIC, y adquirido mediante el espectrómetro de masas Triple TOF ® 5600. (AB Sciex,USA).

1.6. Análisis de los datos: Bases de datos de proteínas y motores de identificación

La mayoría de los repositorios existentes son de acceso público y permiten el acceso tanto por el nombre de la proteína como por la llave primaria o identificador que recibe la entrada cuando se registra en la base de datos. Tanto el nombre de la proteína como su identificador varían dependiendo de la base de datos consultada, lo que obliga a tener siempre presente la referencia de la fuente de datos para determinar con exactitud el ámbito de la información asociada a una determinada entrada.

Aunque existen muchos repositorios, todos bajo la definición genérica de bases de datos de proteínas, no todos almacenan el mismo tipo de información, y varían desde los que almacenan información referente a la secuencia de aminoácidos de la proteína junto a anotaciones funcionales y bibliográficas como SwissProt (http://www.uniprot.org/) o NCBInr (http://www.ncbi.nlm.nih.gov/), a otros que contienen datos relativos a la estructura tridimensional obtenidos por difracción de rayos X, como PDB (http://www.rcsb.org/pdb/home/home.do). No obstante, es requisito indispensable que la base de datos a emplear contenga la secuencia de aminoácidos de las proteínas que recopila.

El otro actor principal en el proceso de identificación de proteínas son los motores de identificación o motores de búsqueda. Estos cumplen dos funciones: primero obtener espectros o patrones teóricos para cada una de las secuencias depositadas en la base de datos, y segundo, proporcionar un método para comparar cada uno de los patrones teóricos con el espectro real obtenido en el espectrómetro. En este segundo aspecto existen varios sistemas. El más simple y obvio consiste en contabilizar las coincidencias entre el número de

péptidos teóricos y reales que se emparejan, en base a lo que se conoce como emparejamiento de picos o *peak matching* para cada una de las secuencias de interés contenidas en la base de datos (Mann, Hojrup et al. 1993; Wilkins, Gasteiger et al. 1998; Clauser, Baker et al. 1999). Sin embargo, este método presenta el inconveniente de supeditar la fiabilidad de la identificación a la longitud de la secuencia comparada (Chamrad, Korting et al. 2004).

La identificación basada en motores de búsqueda la podemos dividir en dos métodos atendiendo al objeto de la identificación. El primer caso es la identificación de proteínas a partir de la huella peptídica (PMF: *Peptide Mass Fingerprinting*). El segundo es la identificación de un determinado péptido a partir de su espectro de fragmentación (PFF: *Peptide Fragment Fingerprinting*). Tanto la huella pepitica como el espectro de fragmentación son espectros de masas con la diferencia que en el primero los picos representados son valores m/z de péptidos, y el segundo, como se ha descrito anteriormente, valores m/z de fragmentos de péptidos La tabla I1.2 resume los motores de búsqueda más empleados tanto para PMF como para PFF.

Nombre	Tipo	URL
MASCOT	PMF,PFF	http://www.matrixscience.com/
ProteinProspector	PMF, PFF	http://prospector.ucsf.edu/prospector/mshome.htm
Profound	PMF	http://prowl.rockefeller.edu/prowl-cgi/profound.exe
ProteinPilot	PFF	http://www.absciex.com/
Sequest	PFF	http://www.thermoscientific.com/
OMSSA	PFF	http://pubchem.ncbi.nlm.nih.gov/omssa/
X!Tandem	PFF	http://www.thegpm.org/TANDEM/index.html
Phenyx	PFF	http://www.genebio.com/
EasyProt	PFF	http://easyprot.unige.ch/
VEMS	PFF	http://portugene.com/vems.html
FindPept	PFF	http://web.expasy.org/findpept/
InsPecT	PFF	http://proteomics.ucsd.edu/Software/Inspect.html
PepFrag	PFF	http://prowl.rockefeller.edu/prowl/pepfrag.html

Tabla I1.2: Motores de búsqueda más empleados en proteómica.

PMF: Huella peptídica (*Peptide Mass Fingerprinting*). PFF: Espectro de identificación (*Peptide Fragment Fingerprinting)*.

En el caso de la huella peptídica, de forma iterativa y a partir de cada una las secuencias de proteínas contenidas en la base de datos, se realizan digestiones virtuales para obtener la relación m/z de cada uno de los péptidos y así poder identificar una determinada proteína. En el caso del espectro de fragmentación, se realizan fragmentaciones teóricas de las secuencias de péptidos para obtener la relación m/z de cada fragmento y así poder identificar un determinado péptido.

A modo de ejemplo, se puede citar el sistema Mascot (www.matrixscience.com) (Perkins, Pappin et al. 1999). Ésta herramienta está basada en el sistema de puntuación estadístico Mowse (Pappin, Hojrup et al. 1993) y permite tanto la identificación de proteínas por huella peptídica como la identificación de péptidos a partir del espectro de fragmentación. MASCOT calcula la probabilidad de un emparejamiento aleatorio entre un pico del espectro y la relación m/z calculada a partir de la secuencia de un péptido o fragmento para cada una de las entradas de la base de datos. Atendiendo al número de picos realmente emparejados y aquellos calculados por azar, se ordenan los candidatos en un ranking de puntuación.

1.7. Proteómica Diferencial Cuantitativa

Gracias al avance de las técnicas analíticas y la mejora en el procesamiento de los datos, la identificación de péptidos y proteínas por espectrometría de masas se puede considerar una tarea prácticamente rutinaria. Esta realidad permite abordar otros aspectos en donde no sólo se determine si una proteína está presente en una muestra, sino los cambios en los niveles de expresión cuando se modifican las condiciones basales (control o sana) o de referencia (tratada o enferma).

En definitiva, se persigue la caracterización de los cambios de la expresión génica a nivel de proteína a partir de la determinación de la abundancia de las mismas, convirtiéndose en una faceta fundamental para la comprensión de los mecanismos regulatorios biológicos (Weston and Hood 2004). Sin embargo, su análisis resulta complejo debido a limitaciones intrínsecas a las técnicas analíticas empleadas y que se traducen en la dificultad de determinar con exactitud la abundancia de las proteínas sólo a partir de los resultados obtenidos por espectrometría de masas.

1.7.1. Análisis de la expresión diferencial de proteínas por electroforesis bidimensional

El uso de geles bidimensionales para determinar la abundancia relativa proteínas en mezclas complejas tiene un largo recorrido (O'Farrell 1975), (Anderson and Anderson 1996). La base es realizar, en el caso más simple, una separación de proteínas por electroforesis bidimensional en geles de poliacrilamida independientes para cada condición y comprobar la variación de intensidad entre los spots de interés. Sin embargo, son muchos los inconvenientes que presenta, debido fundamentalmente a la falta de reproducibilidad de la técnica cuando varias muestras son separadas en geles independientes. Diferentes parámetros experimentales, como por ejemplo los tiempos de desarrollo del agente de tinción, intervienen en la aparición de alteraciones artefactuales, por lo que es difícil realizar una medición exacta del cambio de expresión de los distintos spots. Adicionalmente a estos cambios en la variación de la intensidad, se le añaden los sutiles cambios que los spots presentan en su localización o coordenada dentro del gel-matriz. Para reducir esta variabilidad, una práctica habitual es la elaboración de réplicas técnicas, y por tanto de varios geles, por cada condición a comparar. Esta práctica, aún mejorando el resultado, sigue presentando limitaciones en cuestión de reproducibilidad, exactitud en la medida y sobre todo, capacidad de procesamiento (Monteoliva, Martinez-Lopez et al. 2011).

Una mejora de la electroforesis 2D clásica a la hora de realizar análisis cuantitativos es la técnica DIGE (*DIfference Gel Electrophoresis*) (Unlu, Morgan et al. 1997). Esta reduce las limitaciones anteriormente descritas al combinar el análisis de distintas muestras en un mismo gel-matriz. Para poder diferenciar las distintas muestras o condiciones entre sí, éstas son marcadas al inicio del proceso con fluoróforos distintos pero sin inducir cambios en el comportamiento electroforético de las proteínas presentes (ni en sus cargas aparentes ni en sus masas relativas). El marcaje permite aislar las proteínas provenientes de una determinada condición o muestra. Al utilizar un único gel-matriz, las variaciones artefactuales resultado del empleo de un gel 2D por cada condición disminuyen considerablemente, haciéndose posible extraer un valor cuantitativo que refleja con mayor fiabilidad las variaciones que se establecen entre las diferentes muestras o condiciones analizadas.

Esto se consigue adquiriendo imágenes virtuales a las diferentes longitudes de onda de emisión de los fluoróforos utilizados en el marcaje y, posteriormente, comparando dichas imágenes entre sí. Este valor es un valor relativo de incremento o decremento de una condición o muestra A respecto otra B. En el cálculo de los ratios de interés el gel es escaneado tantas veces como muestras se hayan combinado, a las longitudes de onda de absorción y emisión respectivas de cada fluoróforo. En el empleo de esta técnica se permite, y se recomienda, reservar uno de los fluoróforos para marcar una muestra de referencia. Este estándar interno suele consistir en una mezcla equimolar todas las condiciones experimentales y permite normalizar las intensidades cuando, debido al número de muestras a comparar, es necesario emplear más de un gel.

1.7.2. Cuantificación de proteínas y/o péptidos por espectrometría de masas

Mientras que las técnicas analíticas para la identificación masiva de proteínas por espectrometría de masas están bien establecidas, no ocurre lo mismo con las destinadas a la denominada proteómica diferencial cuantitativa. Uno de los inconvenientes al utilizar directamente los espectros de masas para determinar la abundancia de ciertas proteínas a lo largo de una serie de muestras o condiciones es que la espectrometría de masas no es intrínsecamente cuantitativa para el análisis de los péptidos. Esto se debe a que la señal correspondiente a un determinado péptido es función no sólo de su cantidad relativa sino también de sus propiedades físico-químicas que afectan a la eficiencia de ionización.

Para llevar a cabo un estudio de cuantificación diferencial basado en espectrometría de masas, son varias las técnicas disponibles (Ong and Mann 2005) (Fig. I1.8). Estas se pueden agrupar en: 1) el marcaje isotópico diferencial de las muestras a comparar, que puede ser in vivo o in vitro, mediante el empleo de compuestos o reactivos que se incorporan al proteoma a analizar y permiten reconocer en todo momento las muestras que están siendo comparadas. Las variaciones de intensidad de cada péptido identificado, detectables en el espectro de masas, pueden relacionarse directamente con la variación de abundancia de las proteínas entre las diferentes muestras; 2) la inclusión de estándares internos en las muestras, también isotópicamente distinguibles, permiten realizar un análisis por comparación de ratios de variación contra el estándar[6]; estas aproximaciones se conocen como Monitorización Selectiva de iones o SRM (*Selected Reaction Monitoring*) y 3) análisis bioinformáticos de datos de espectrometría de masas sin tratamiento previo, llamadas estrategias libres de marca o *label-free*.

Técnicas de marcaje diferencial

En proteómica diferencial cuantitativa el marcaje diferencial puede ser metabólico o químico. En el caso metabólico, el abordaje SILAC (*Stable Isotope Labelling by Amino Acids in Cell Culture*) (Ong, Blagoev et al. 2002) se basa en el marcaje durante el cultivo de una línea celular crecida en dos condiciones diferentes donde en una de las condiciones uno o más aminoácidos del medio de cultivo son sustituidos por versiones marcadas con isótopos pesados estables (por ejemplo ^{13}C en lisina (K) o arginina (R)). El uso de distintos isótopos para el marcaje hará que dos péptidos con la misma secuencia proporcionen dos valores diferentes de masa molecular y por tanto relación m/z. Esta diferencia varía según el marcaje utilizado y recibe el nombre de desplazamiento de masa o *mass shift* y permite alinear y comparar los péptidos de la misma

6 Siempre y cuando el estándar sea idéntico o químicamente idéntico.

proteína obtenida en dos condiciones diferentes. Estos péptidos, no obstante, serán indistinguibles de sus equivalentes ligeros en el resto de sus propiedades físico-químicas incluyendo su comportamiento cromatográfico. La intensidad relativa de las señales ligera y pesada de un mismo péptido permite estimar la diferencia de concentración de dicho péptido y, por extensión, de su proteína de origen en las dos condiciones experimentales.

Una vez analizada la mezcla de péptidos por espectrometría de masas, los elementos derivados de la señal utilizados para calcular la abundancia de los péptidos identificados en la diferentes muestras, denominados *features*, son los picos MS^1 o relación m/z de los iones precursores de los péptidos. Para poder calcular la variación relativa de intensidad entre las versiones ligera y pesada del mismo péptido, basta localizar un par de picos MS^1 asociados al mismo péptido y comprobar que experimentan una diferencia de relación m/z que consistente con el desplazamiento de masa introducido. La diferencia entre las intensidades de ambas señales MS^1 proporcionará la variación relativa entre ambos péptidos. Finalmente y una vez localizados e identificados las versiones ligera y/o pesada de todos los péptidos que pertenecen a una determinada proteína, se calculará el valor promedio de la variación relativa de abundancia de la proteína en cuestión.

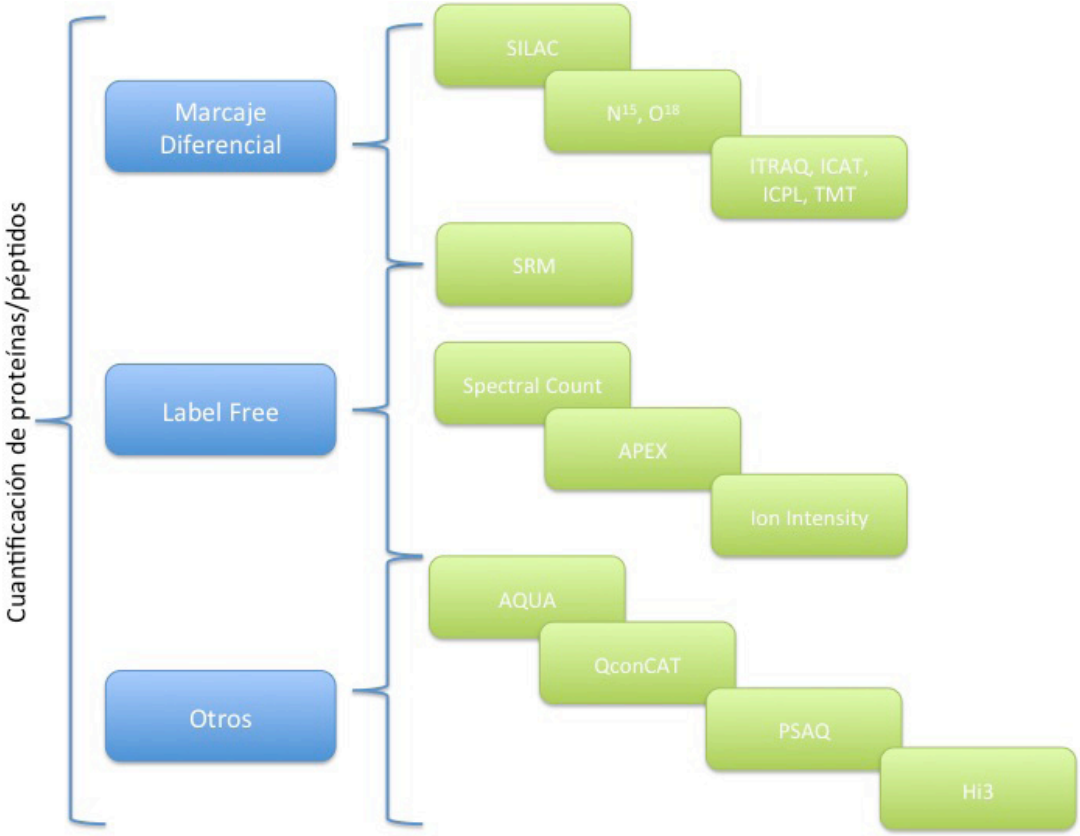

Figura I1.8: .Estratégias proteómica cuantitativa. Adaptación de (Gonzalez-Galarza, Lawless et al. 2012)

Otras técnicas de marcaje diferencial ampliamente utilizadas son ICAT (*isotope-coded affinity tag*) (Gygi, Rist et al. 1999), ICPL (*isotope-coded protein label*) (Schmidt, Bisle et al. 2009), TMT (*tandem mass tags*) (Thompson, Schafer et al. 2003) y iTRAQ (*isobaric tag for relative and absolute quantitation*) (Wiese, Reidegeld et al. 2007) que se introducen por marcaje químico bien a nivel de péptido o de proteína. El marcaje mediante iTRAQ o TMT se diferencian del resto en que la comparación de las abundancias de los péptidos se realiza en los espectros de fragmentación, donde se comparan las intensidades de los iones etiqueta (*reporter ions*). En el caso del sistema *4-plex*® de iTRAQ, los iones etiqueta tienen valores m/z 114,115,116 y 117 Da lo que permite calcular la variación en abundancia de cada péptido hasta en cuatro muestras o condiciones distintas.

Monitorización selectiva de iones (SRM)

Aunque esta técnica ha sido incorporada recientemente en proteómica cuantitativa (Kuzyk, Smith et al. 2009; Malmstrom, Beck et al. 2009; Picotti, Bodenmiller et al. 2009), su utilización para la cuantificación de pequeñas moléculas como metabolitos cuenta con un recorrido más amplio (Murray and Watson 1986). La implementación del análisis SRM requiere el uso de un tipo específico de espectrómetro de masas que incorpore un analizador de tipo triple cuadrupolo (Fig I1.9). Este instrumento permite un doble filtrado de iones provenientes de la misma molécula. En un primer paso, el primer cuadrupolo o Q1, permite el aislamiento de un péptido en base a la relación m/z de su ion precursor. En el segundo cuadrupolo o Q2, el péptido es fragmentado generando iones fragmento. Por último, el tercer cuadrupolo o Q3, filtra un único ión fragmento que es detectado a continuación. A la combinación de de los valores de m/z filtrados en Q1 y Q3 se le denomina transición. Una proteína es detectada a partir de varios péptidos que, a su vez, son monitorizados a partir de varias transiciones. El hecho de que el SRM sea una técnica dirigida (es necesario decidir *a priori* que proteínas queremos analizar) combinado con el uso de transiciones que son monitorizadas a muy alta velocidad le confiere una alta especificidad, sensibilidad y amplio rango dinámico de cuantificación (Anderson and Hunter 2006; Keshishian, Addona et al. 2007; Stahl-Zeng, Lange et al. 2007).

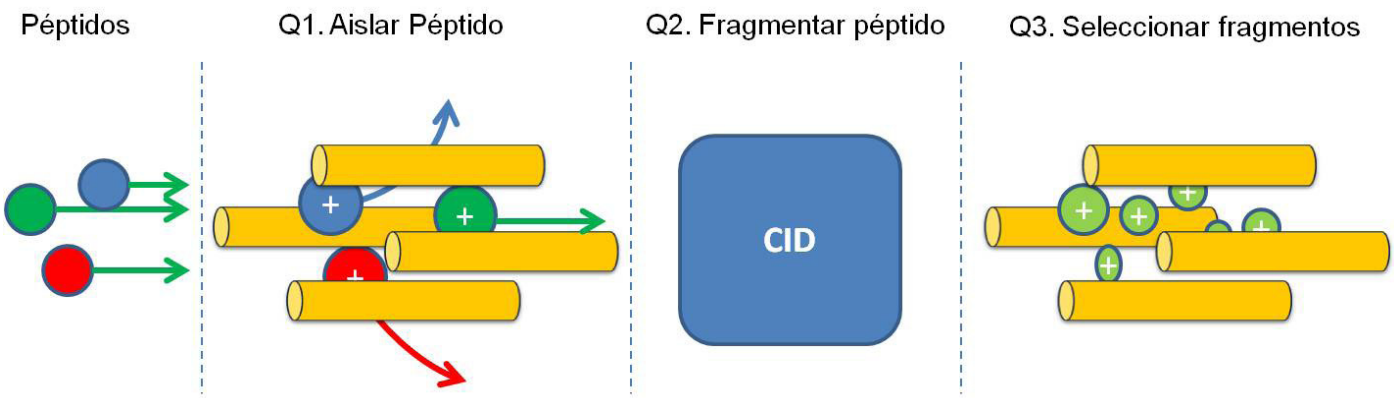

Figura I1.9: Esquema triple quadrupolo.

Aproximaciones libres de marca o marca Label-free

Las técnicas denominadas sin marcaje o *label-free* son métodos cuya principales características son su bajo coste y su relativa simplicidad aunque requiere un sistema cromatográfico extremadamente preciso que asegure la reproducibilidad experimental. Estos métodos pueden agruparse en dos tipos:

1) Aquellos que realizan una medición de la intensidad de la señal MS de los iones precursores (MS^1) en análisis por LC-MS consecutivos.

2) Aquellos que contabilizan el número de veces que cada péptido es fragmentado e identificado (por acumulación de espectros o *spectral counting*). Esta técnica se basa en la asunción de que los péptidos más abundantes serán fragmentados e identificados con una mayor frecuencia que los menos abundantes.

Estándares de datos proteómicos

2.1. Comunicación entre datos proteómicos

El punto anterior introduce las principales técnicas analíticas utilizadas en proteómica para separar, identificar y cuantificar proteínas presentes en proteomas complejos. Sin embargo, este progreso experimental no ha venido acompañado de un compromiso global de los formatos y esquemas para representar los datos asociados a los resultados y que potenciarían un intercambio real de información entre la comunidad académica así como una verificación y validación independiente.

Teniendo como referente un esquema clásico de un experimento proteómico, donde se sigue de manera ordenada las siguientes fases: extracción y separación de proteínas, análisis mediante espectrometría de masas, identificación de péptidos y proteínas y cuando el estudio así lo requiera y permita, la cuantificación de las mismas, los datos generados en cada una de las fases son dependientes de la aproximación analítica empleada así como del instrumento y paquetes software utilizados. Esto implica una serie de limitaciones directamente relacionadas con la independencia de los experimentos, al no poder desvincular el resultado de los mismos de su origen analítico o con la transparencia del procesamiento al que son sometidos. Los paquetes de software, la mayoría de ellos de origen comercial, no permiten un conocimiento exhaustivo de cómo se lleva a cabo este procesamiento. Otro factor limitante que resulta especialmente crítico reside en el intercambio de información en aquellos proyectos que por su envergadura son inabordables desde una única plataforma tecnológica.

2.2. HUPO PSI

La iniciativa de Estándares en Proteómica (*Proteomics Standards Initiative* - PSI*)*, promovida por la Organización Internacional del Proteoma Humano (*Human Proteome Organisation* - HUPO), lleva más de diez años liderando los progresos y mejoras en relación a la estandarización de los datos obtenidos en experimentos proteómicos. Una de las principales características y clave del éxito de este grupo es la implicación concurrente tanto del mundo académico como del industrial, alcanzando una colaboración abierta hacía la comunidad científica generando y/o proponiendo por consenso estándares de comunicación

Tal y como su sitio web describe (www.hupo.org/research/psi/), el principal objetivo de esta iniciativa es***: "Definir los estándares para la representación de datos en proteómica que facilite la comparación, intercambio y verificación de estos datos***". (Kaiser 2002; Orchard, Hermjakob et al. 2003). Las bases para la constitución PSI fueron los trabajos desarrollados para la definición del primer repositorio público que almacenaba datos experimentales proteómicos y aquellos que proponían los primeros formatos para almacenar e intercambiar información analítica independientemente a la plataforma de espectrometría de masas empleadada en los laboratorios de proteómica (Taylor, Paton et al. 2003; Carr, Aebersold et al. 2004; Garwood, McLaughlin et al. 2004; Pedrioli, Eng et al. 2004).

Desde su fundación en 2002, los grupos que han colaborado con la iniciativa han organizado distintas conferencias, eventos, reuniones de trabajo y publicaciones que han contribuido de manera relevante en la consecución de los objetivos (globales y particulares) (Orchard, Kersey et al. 2003; Orchard, Kersey et al. 2003; Orchard, Zhu et al. 2003; Orchard, Hermjakob et al. 2005; Orchard, Hermjakob et al. 2005; Orchard, Apweiler et al. 2006; Orchard, Hermjakob et al. 2006; Orchard, Montechi-Palazzi et al. 2007; Orchard, Taylor et al. 2007; Orchard, Albar et al. 2008; Orchard, Martens et al. 2008; Orchard, Deutsch et al. 2009; Orchard, Hoogland et al. 2009; Orchard, Albar et al. 2010; Orchard, Jones et al. 2010; Orchard, Albar et al. 2011; Orchard, Albar et al. 2012). Para la distribución de las tareas concretas asociadas a los objetivos marcados, PSI se estructura en los siguientes grupos de trabajo:

a) Interacciones moleculares *(Molecular Interactions*).

b) Espectrometría de masas (*Mass Spectrometry*).

c) Procesamiento computacional de datos de MS (*Proteomics Informatics*).

d) Modificaciones de proteínas (*Protein Modifications*)

e) Separación de proteínas (*Protein Separation*).

Los objetivos comunes que una vez materializados se aplican de manera individualizada para cada grupo de trabajo son:

1) Elaboración de la guías que describan la información mínima necesaria para compartir o publicar un experimento proteómico.

2) Definición de formatos de intercambio formales que permitan la comunicación entre paquetes software o para poder enviar los resultados a bases de datos accesibles públicamente.

3) Vocabularios controlados que permitan una terminología estándar para describir los elementos contenidos en los formatos de datos.

4) Soporte para la implementación de los estándares en las herramientas públicas disponibles.

2.3. Información de un experimento: MIAPE y estándares XML

Dentro del primero de los objetivos marcados por la iniciativa, el PSI publicó en 2007, la especificación MIAPE (*Minimum Information About a Proteomics Experiment)* (Taylor, Paton et al. 2007). A partir de este documento troncal se desarrollaron una serie de módulos MIAPE (tabla I2.1) para las distintas áreas de mayor interés dentro del campo de la proteómica. Más en detalle, un documento MIAPE se compone de la lista mínima de elementos que tras completarse y validarse permite reproducir un experimento. Los elementos que componen un documento MIAPE se rellenan utilizando lenguaje natural y describen puntos concretos tanto del protocolo experimental como del análisis efectuado que permita a otros grupos y usuarios una interpretación sin ambigüedad de los resultados obtenidos a la vez que permitir su replicación. Distintas revistas del campo proteómico (Orchard, Binz et al. 2009; Orchard and Ping 2009) requieren a los autores cumplir con estas recomendaciones para asegurar la reproducibilidad de los datos.

De forma simultánea a la elaboración de los guías MIAPE, HUPO PSI definió una serie de formatos de intercambio basados en el estándar XML (*Extensible Markup Language)* (tabla I2.1) para permitir una completa integración de los datos independientemente de su origen y/o software utilizado para su procesamiento.

En el contexto de un experimento típico basado en espectrometría de masas, son dos los principales formatos disponibles. El primero de éstos es el mzML (versión 1.1.0) (Martens, Chambers et al. 2011) y comprende los datos los relativos a la adquisición realizada por el espectrómetro de masas. La información se puede incorporar tanto como lista de picos sin procesamiento alguno, meramente un muestreo de la señal que el espectrómetro genera, como procesadas mediante filtros, suavizados o la aplicación de umbrales según la relación señal-ruido (SNR, *Signal to Noise Ratio*) y preparadas para ser empleadas en la posterior identificación. En relación a esta fase de identificación, el estándar seguido es el mzIdentML (versión 1.1.0) (Jones, Eisenacher et al. 2012) y se encarga de capturar la información que se desprende de los motores de búsqueda con la identificación de péptidos y proteínas.

De forma complementaria, y una vez formateados los resultados según mzML y mzIdentML, en el caso de realizar un experimento cuantitativo, son dos los estándares que pueden integrar esta información, mzQuantML (Walzer, Qi et al. 2013) y mzTab (https://code.google.com/p/mztab/). El primero realiza una captura detallada de los datos reportados, mientras que el segundo reúne un resumen de los resultados.

Por último, en el caso de que las proteínas se hayan obtenido mediante una separación mediante el uso de electroforesis el formato estándar es el GelML (Gibson, Hoogland et al. 2010).

	Guias MIAPE	Formato de Intercambio
Espectrometría de masas	Mass Spectrometry (MIAPE-MS) (Taylor, Binz, et al. 2008)	mzData (obsoleto)
		mzML (Martens, Chambers et al. 2011) (ver. 1.1.0 estable) **http://www.psidev.info/mzml**
		TraML (Deutsch, Chambers et al. 2012) (ver.1.0.0 estable) **http://www.psidev.info/traml**
	Mass Spectrometry Informatics (MIAPE-MSI) (Binz, Barkovich et al. 2008)	mzIdentML (Jones, Eisenacher et al. 2012) (ver.1.1.0 estable) **http://www.psidev.info/mzidentml**
	Quantitation (MIAPE-Quant) (Martinez-Bartolome, Deutsch et al. 2013)	mzQuantML (Walzer, Qi et al. 2013) (ver. 1.0.0 estable) **http://www.psidev.info/mzquantml** mzTab [enviado] *(ver.1.0.0 candidata)* **http://psidev.info/mztab**
Interacciones moleculares	Interactions (MIMIx) (Orchard, Salwinski et al. 2007)	PSI-MI XML (ver.2.5.4 estable) **http://www.psidev.info/mif** MiTAB (ver. 2.5 estable) **http://wiki.reactome.org/index.php/PSI-MITAB_interactions**
Separación de proteínas	Gel Electrophoresis (MIAPE-GE) (Gibson, Anderson et al. 2008)	GelML (Gibson, Hoogland et al. 2010) (v.1.1.0 estable) **http://www.psidev.info/gelml**
	Gel Informatics (MIAPE-GI) (Hoogland, O'Gorman et al. 2010)	Ninguno
Procesamiento de la muestra	Column chromatography (MIAPE-CC) (Jones, Carroll et al. 2010)	spML (obsoleto)
	Capillary electrophoresis (MIAPE-CE) (Domann, Akashi et al. 2010)	

Tabla I2.1: Guias MIAPE y formatos PSI XML de intercambio de información desarrollados por HUPO.PSI. Adaptación de (Medina-Aunon, Krishna et al. 2013).

2.4. Estándar para experimentos basados en espectrometría de masas: PRIDE XML

Aunque los estándares anteriormente descritos son actualmente una realidad en el seno de muchos laboratorios de proteómica, este hito no se ha materializado hasta una época más reciente. En esta transición, los esfuerzos estuvieron dirigidos a diseñar un formato que capturase sólo una porción de los datos incluidos en los formatos estándares que aunque no asegurasen la reproducibilidad de los datos si permitiese el intercambio de los resultados basados en espectrometría de masas. Este formato se denominó PRIDE XML.

Este formato no pretendía rivalizar con los formatos XML propuestos por el PSI. Se define un esquema sencillo que, alternativamente a los primeros borradores del estándar mzIdentML sobre identificación de péptidos y proteínas, pudiese utilizarse para la recogida, almacenaje y consulta por parte de la comunidad científica de este tipo de información en un repositorio centralizado. Este repositorio recibió el nombre de PRIDE (*PRoteomics IDEntification databse*) (Martens, Hermjakob et al. 2005; Jones, Cote et al. 2006; Csordas, Ovelleiro et al. 2012) y cuenta en la actualidad con más de 28000[7] experimentos basados en espectrometría de masas, convirtiéndose de facto en el principal repositorio público en esta materia.

7 Dato proveniente del sitio web PRIDE (http://www.ebi.ac.uk/pride). 28431 experimentos a 16 de Julio de 2013.

Extracción de información biológica

Uno de los hitos que se busca en cualquier experimento en donde intervienen técnicas de análisis de alto rendimiento como genómica y proteómica es determinar el papel biológico de los genes expresados o las proteínas identificadas, situándolas en el contexto experimental de donde provienen. En los últimos años y gracias al desarrollo exponencial de bases de datos y repositorios públicos (Buckingham 2004; Martens 2011) es posible acceder en tiempo real a la información biológica y funcional disponible a partir de una serie de consultas sencillas.

Son dos las estrategias que tradicionalmente se han aplicado. La primera es la minería de texto o *text mining* (Hearst 1999) cuyo objetivo es examinar un conjunto de documentos escritos en lenguaje natural, y por tanto no estructurados en un lenguaje formal, para obtener una relación entre los mismos sin ninguna premisa inicial (Nasukawa and Nagano 2001) que puede o no resolver la pregunta de un sujeto. El objetivo es determinar la relación entre los elementos o documentos que forman el conjunto de trabajo, más que la adecuación de la respuesta a la pregunta a resolver.

La segunda es la denominada recuperación de información o *information retrieval* (Cowie and Lehnert 1996). Está basada en el almacenamiento, organización y acceso a la información. Al tratarse de información estructurada y etiquetada en muchos casos, en lugar de analizar un conjunto de textos y extraer sus posibles vínculos, se sabe a priori la relación que hay entre los elementos que forman el sistema de búsqueda y por tanto si la relación se corresponde con la pregunta que realiza el individuo. Al contrario que la minería de texto, el objetivo de este enfoque si es adecuar la respuesta lo más exactamente posible a la pregunta inicial. A continuación se exponen algunos de los principales recursos disponibles a través de la web que permiten la extracción de conocimiento biológico.

3.1. Literatura científica

Una de las fuentes más valiosas son los artículos, manuscritos y otras referencias bibliográficas que versan sobre los ensayos y experimentos científicos. En este sentido PubMed (www.ncbi.nlm.nih.gov/pubmed) es el principal recurso y cuenta con más de 23 millones de citas bibliográficas indexadas por MEDLINE (Delwiche 2008).

Una búsqueda típica de fuentes bibliográficas se realiza a partir de una serie de consultas booleanas (atributo igual a valor) sencillas que pueden combinarse gracias a los operadores lógicos Y (intersección de conjuntos) y O (unión de conjuntos). El sistema de búsquedas se complementa con un historial, para poder combinar consultas realizadas y también con un índice de referencias relacionadas, que permite al usuario realizar una exploración por los distintos temas que se visualizan con la idea de acotar lo mejor posible el ámbito de su consulta.

A pesar de ser fácil e intuitiva, este tipo de interfaces presentan los problemas típicos de procesamiento del texto como la falta de uniformidad en la redacción de los artículos, requiriendo por tanto un análisis en detalle de los textos obtenidos para poder extraer un conocimiento más focalizado en el área de interés.

3.2. Bases de datos de anotaciones

La recuperación de la información está directamente relacionada con las bases de datos y repositorios de anotaciones. A pesar de que en el punto 2.4 ya se introducía el repositorio PRIDE, como principal sitio donde depositar los resultados experimentales basados en espectrometría de masas según el formato estándar PRIDE XML, los sistemas resumidos en la tabla I3.1 son algunos de más utilizados que permiten conocer las anotaciones relacionadas con una determinada proteína.

TIPO DE DATO	URL	INFORMACIÓN
Características	http://www.uniprot.org/	Nombre de proteína y gen, secuencia de aminoácidos, alias e identificadores alternativos.
Anotaciones	http://www.uniprot.org/	Función, localización celular, especificidad de tejido, referencias patológicas.
Palabras clave	http://www.uniprot.org/	Lista de términos controlados que resume las principales características de cada entrada.
Taxonomía	http://www.ncbi.nlm.nih.gov/Taxonomy/	Ontología de taxonomías
Gene Ontology		
Amigo	http://amigo.geneontology.org/	Interfaz de búsqueda de términos de ontología génica GO.
Rutas metabólicas		
KEGG	http://www.genome.jp/kegg/pathway.html	Colección de mapas y rutas metabólicas.
Interacciones de proteínas		
STRING	http://string-db.org/	Interacciones conocidas y predichas.
IntAct	http://www.ebi.ac.uk/intact/	Base de datos y herramientas de análisis para los datos derivados de interacciones de proteínas.

TIPO DE DATO	URL	INFORMACIÓN
Respositorios experimentales		
HPRD	http://www.hprd.org/	Visualización e integración de la información relativa al proteínas humanas.
Modificaciones postraducionales		
Phosphosite	http://www.phosphosite.org/	Base de datos de fosforilzaciones.
Dominios funcionales y familias de proteínas		
Interpro	http://www.ebi.ac.uk/interpro/	Base de datos utilizada par la clasificación y anotación automática de proteínas.
Otras		
PharmaGKB	http://www.pharmgkb.org/	Base de datos de fármacos y su enlace con estudios genéticos y genómicos.

HPRD, *human protein reference database*.

STRING, *Search Tool for the Retrieval of Interacting Genes/Protein.*

Tabla I3.1: Listado de direcciones y recursos computacionales disponibles para la recuperación de anotaciones de proteínas (Medina-Aunon, Paradela et al. 2010).

3.3. Vocabularios controlados y ontologías

Los vocabularios controlados y las ontologías son elementos críticos a la hora de recuperar la información asociada a una determinada proteína. Ambos presentan diferencias considerables: mientras que los vocabularios controlados pueden equipararse con simples enumeraciones de palabras agrupadas bajo un determinado concepto, las ontologías presentan una estructura más compleja basada en una jerarquía de conceptos que permite establecer distintos tipos de relaciones entre estos. Un ejemplo de vocabulario controlado sería aquel que incluye unidades de medidas, como centímetros, metros, kilómetros, litros, hectáreas y en el caso de las ontologías, además de incluir todos los elementos anteriores, se podrían definir relaciones según el objeto de medición: lineal (centímetro, metro, kilómetro), superficie (hectáreas) o volumen (litro). En una aplicación práctica, el vocabulario definido puede limitar los valores posibles que pueda tomar un determinado atributo o variable (tiene que ser una de las medidas enumeradas), mientras que la ontología permite además de hacer referencia y limitar a una categoría, como distancia lineal, incluir la semántica del atributo.

Ambas estructuras permiten centrarse en la búsqueda y posterior análisis de los resultados en palabras, términos y conceptos concretos que permitan una extracción de la información más exacta. En lugar de analizar detalladamente cada anotación, se emplean sistemas de minería de datos para determinar la semántica de éstos. Así, evitamos en primer término la ocurrencia de errores tipográficos, las palabras que se incluyen en las anotaciones son validadas en el momento de incorporarlas, y en segundo la combinación de conceptos que no presenten ninguna relación. Los vocabularios controlados más frecuentes en la recuperación de anotaciones son los que se presentan en la tabla I3.2 siendo algunos de ellos mantenidos por el HUPO PSI dentro de su labor de estandarización.

ONTOLOGÍA	URL	INFORMACIÓN
BRENDA	http://www.brenda-enzymes.info/	Tejidos, líneas celulares y tipos celulares.
PSI MS	http://www.obofoundry.org/	Espectrometría de masas.
PSI MI	http://www.obofoundry.org/	Interacciones proteína-proteína.
PSI MOD	http://www.obofoundry.org/	Modificaciones de proteínas.
NEWT	http://www.uniprot.org/taxonomy/	Organización de taxonomías y organismos.
PW	http://www.obofoundry.org/	Rutas metabólicas.
UO	http://www.obofoundry.org/	Unidades de medida.
SEP	http://www.obofoundry.org/	Procesamiento de la muestra y técnicas de sepración.

Tabla I3.2: Principales ontologías utilizadas en proteómica.

3.4. Identificadores y formatos de salida

Tal y como se describía en la sección 1.6, otros elementos a analizar y resolver a la hora de recuperar las anotaciones relacionadas con una determinada proteína son la base de datos que contiene la entrada y el tipo de identificador que se asocia a ésta. Debido a que son muchas y diversas las bases de datos que son utilizadas para la identificación de proteínas, la forma de construir los identificadores de cada proteína varía, siendo necesario conocer siempre cual es la base de datos de origen para poder determinar con exactitud la entrada que está siendo utilizada.

De forma análoga a la dependencia que existe en las bases de datos en relación al formato de entrada, los formatos de salida que presentan la información asociada a una determinada proteína varían según la fuente consultada. Además del tipo de salida que varía por lo regular entre salidas editables como son texto plano, HTML o XML, hay que analizar la estructura de cada una para determinar con exactitud donde se encuentra la anotación de interés. En algunos casos, gracias a los vocabularios controlados y ontologías, este trabajo resulta más asumible, sin embargo existen otras ocasiones que debido a una falta de estructura y organización del texto, requiere una búsqueda más compleja y difícilmente abordable de forma automática y desentendida.

3.5. Sistemas de recuperación en lotes

Teniendo en cuenta las limitaciones mencionadas en el apartado anterior para llevar a cabo una recuperación exacta de la información, el proceso a realizar es tedioso y difícilmente manejable. Cabe destacar que es necesaria cierta cautela en la selección de la base de datos fuente para obtener una mejor comprensión del contexto de los resultados y las tareas relacionadas con la minería de datos para retener lo realmente relevante desde el punto de vista del objetivo del experimento. Junto con estos inconvenientes y pese a que el proceso puede ser modelado, sin el correspondiente apoyo de sistemas que permitan implementar dicho modelo en un entorno computacional, el proceso sería sólo abordable para un reducido conjunto de proteínas y no para resultados reales donde intervienen cientos o miles de proteínas.

Para solventar estos inconvenientes, se han desarrollado diferentes herramientas en los últimos años. Las primeras de estas se enmarcaban en un contexto puramente genómico. En este sentido, se puede considerar como la primera herramienta en el campo la desarrollada por Khatri y sus colaboradores (Khatri, Draghici et al. 2002) que almacenaba todos los términos de la ontología Gene Ontology (GO) (Ashburner, Ball et al. 2000) (www.geneontology.org) en una base de datos local y asociaba cada uno de estos con la función bioquímica y la localización en el cromosoma para un conjunto de genes. A continuación una lista de genes problema se enfrentaba contra la información almacenada en esta base de datos, resultando como salida en primer término del sistema aquellas anotaciones para cada uno de los genes de entrada que estaban registradas en la base de datos, para, a continuación, clasificar, mediante técnicas de agrupamiento o *clustering,* los genes según las relaciones derivadas de la ontología GO.

Posteriormente, GO ha servido como base principal para agrupar un conjunto de genes. GoMiner (Zeeberg, Feng et al. 2003) DAVID (*Database for Annotation, Visualization and Integrated Discovery*) (Dennis, Sherman et al. 2003), GOTree Machine (Zhang, Schmoyer et al. 2004) and GOToolBox (Martin, Brun et al. 2004) son algunas de las herramientas que se apoyan en bases de datos locales donde se almacenan la ontología GO junto a anotaciones funcionales como rutas metabólicas o procesos patológicos.

Sin embargo el principal inconveniente de las herramientas listadas anteriormente es su utilización en un entorno proteómico. Sólo por enunciar algunas de las principales limitaciones, es destacable el reducido número de tipos de identificadores que soportan, normalmente sólo uno, y también el enfoque genómico de análisis completo frente al proteómico. Más en detalle, cuando un experimento genómico es analizado se tiene una visión completa de todos los genes que comprenden el experimento, obteniendo un valor cuantificable según el nivel de expresión para cada uno de estos genes. En el caso de proteómica, aquellas proteínas que no se han detectado no aportan ningún valor relacionado con su nivel de expresión, y por tanto limitan su interpretación, al no poder determinar si realmente se trata de un caso de falta de expresión o limitaciones relacionadas con la sensibilidad de detección en la fase analítica. Esto hecho reduce el análisis sólo a aquellas proteínas que si se han podido detectar y en el mejor de los casos identificar.

En un contexto mixto genómico-proteómico, algunas herramientas que se encuentran son FatiGO (Al-Shahrour, Diaz-Uriarte et al. 2004) BABELOMICS (Al-Shahrour, Minguez et al. 2005) WebGestalt (Zhang, Kirov et al. 2005) y GENECODIS (Carmona-Saez, Chagoyen et al. 2007). Todas estas proporcionan las relaciones entre los genes y proteínas que se pueden establecer partiendo de la ontología GO, incluyendo además otras fuentes como son las rutas metabólicas comprendidas en la Enciclopedia de Genes de Kyoto (KEGG *Kyoto Encyclopedia de Genes*) (Kanehisa and Goto 2000) o los términos controlados de la base de datos UniProt/SwissProt (*Keywords*) en relación al papel funcional y estructural, localización celular o la técnica análisis empleada para cada una de las proteínas incluidas en la lista inicial.

Publicaciones: Aportación específica del autor

La siguiente sección enumera los trabajos compendiados en esta tesis doctoral en dos apartados. En primer lugar y por el orden de citación establecido en el capítulo de resultados, se describen las publicaciones reunidas troncalmente en el cuerpo de estos. En segundo lugar, se enumeran otras publicaciones citadas a lo largo de los distintos capítulos de esta memoria. Independientemente del apartado, para todos los trabajos reunidos se incluye una breve descripción de la aportación específica del autor, aunque aquellos que han sido catalogados como troncales y que son descritos en profundidad en la sección de resultados, presentan una mayor extensión y detalle del rol que el autor desempeñó. En aquellos trabajos que se incluye desarrollo de software, en todos los casos este fue creado siguiendo un modelo en cascada: requerimientos, análisis, diseño, implementación e implantación, verificación, validación y pruebas.

4.1. Publicaciones troncales/resultados

1. The gel electrophoresis markup language (GelML) from the Proteomics Standards Initiative*. Gibson F, Hoogland C, Martinez-Bartolomé S, **Medina-Aunon JA,** Albar JP, Babnigg G, Wipat A, Hermjakob H, Almeida JS, Stanislaus R, Paton NW, Jones AR.
Proteomics. 2010 Sep;10(17):3073-81. doi: 10.1002/pmic.201000120.
Índice de impacto 4.132/4.223 (Q1)[a]. Cat. CO. Rank. 14/74

A partir de los casos de uso y posibles escenarios de aplicación de esta técnica de separación, *Medina-Aunon* formó parte del grupo de trabajo de HUPO-PSI (*HUman Proteome Organization - Proteomics Standards Initiative*) encargado de la definición de un modelo lo suficientemente genérico para poder agrupar en primer

término todos los escenarios posibles y en segundo lugar, detallado, para recoger toda la información relevante que permitiese la reproducción de un experimento de este tipo. Una vez que el modelo estuvo lo suficientemente validado y consensuado, fue sometido al proceso de revisión de HUPO-PSI para su categorización como estándar. En paralelo a este proceso, *Medina-Aunon* fue el responsable del diseño e implementación del software que permitía la exportación de datos electroforéticos en este nuevo formato. Este software se basada en el modelo de anotación de experimentos basados en electroforesis recogido en las guías MIAPE (*Minimun Information About a Proteomics Experiment*) y mapeaba las secciones de estas guías con el diseño del estándar. Finalmente participó en la redacción del artículo*.

2. A DIGE study on the effects of salbutamol on the rat muscle proteome - an exemplar of best practice for data sharing in proteomics. Medina-Aunon JA, Kenyani J[1], Martinez-Bartolomé S, Albar JP, Wastling JM, Jones AR. *BMC Res Notes*. 2011 Mar 28;4:86. doi: 10.1186/1756-0500-4-86.
Índice de impacto 1.39

Debido a las limitaciones que el formato PRIDE XML presentaba para poder capturar la información relativa a la separación por electroforesis, *Medina-Aunon* lideró y fue el principal impulsor, diseñador y único desarrollador de la aplicación PRIDESpotMapper para corregir dicha limitación. Para ello, y debido a la experiencia adquirida en la definición del estándar GelML, estudió la forma en como poder capturar los datos necesarios desde los ficheros fuente de espectrometría de masas e identificación de peptídos y proteínas y la forma en como incorporar estos datos en un fichero de resultados PRIDE XML sin alterar el esquema de éste. Igualmente participó en la redacción del artículo. Actualmente es el encargado de su mantenimiento.

3. The ProteoRed MIAPE web toolkit: a user-friendly framework to connect and share proteomics standards. Medina-Aunon JA, Martínez-Bartolomé S, López-García MA, Salazar E, Navajas R, Jones AR, Paradela A, Albar JP. *Mol Cell Proteomics.* 2011 Oct;10(10):M111.008334. doi: 10.1074/mcp.M111.008334.
Índice de impacto 7.251/8.051 (Q1)[a]. Cat. CO. Rank. 5/74

Los estándares de datos son esenciales para poder compartir los resultados de una forma normalizada, favoreciendo un posterior reanálisis por una plataforma analítica distinta a la empleada originalmente. Los distintos formatos XML y guías MIAPE propuestos por HUPO PSI en unión con el formato PRIDE XML son el punto inicial para lograr una colaboración más transparente y fructífera entre la comunidad científica. Con el fin de acercar estos estándares y guías al flujo cotidiano de un laboratorio de proteómica, *Medina-Aunon*, lideró la concepción e implementación del entorno de trabajo ProteoRed MIAPE Web Toolkit (PMWTK) en donde los distintos formatos fuesen traducidos a modelo de objetos que permitiesen una conexión automática y completa. De forma paralela, se mapeo este modelo de objetos al modelo relacional de la base de datos MIAPE (publicación número 9). Además del papel desempeñado en el diseño fue el desarrollador todo lo relativo al intercambio e integración de información según el formato GelML, incluyendo toda la fase de validación que se describe en los resultados, y el principal impulsor de la traducción y mapeo del modelo relacional al formato PRIDE XML. Lideró la redacción y la posterior revisión del mismo una vez obtenidos las objeciones de los revisores.

1 Autoría compartida

4. The mzQuantML Data Standard for Mass Spectrometry-based Quantitative Studies in Proteomics*. Walzer M, Qi D, Mayer G, Uszkoreit J, Eisenacher M, Sachsenberg T, Gonzalez-Galarza FF, Fan J, Bessant C, Deutsch EW, Reisinger F, Vizcaíno JA, **Medina-Aunon JA**, Albar JP, Kohlbacher O, Jones AR. *Mol Cell Proteomics*. 2013 Aug;12(8):2332-40. doi: 10.1074/mcp.O113.028506.
Índice de impacto 7.251/8.051 (Q1)[a]. Cat. CO. Rank. 5/74.

A partir de los casos de uso y posibles técnicas de proteómica cuantitativa basada en espectrometría de masas, *Medina-Aunon* formó parte del grupo de trabajo de HUPO-PSI encargado de la definición de un modelo lo suficientemente genérico para poder agrupar en primer término todos los escenarios posibles y en segundo lugar, detallado, para recoger toda la información relevante que permitiese la reproducción de un experimento de este tipo. Una vez que el modelo estuvo lo suficientemente validado y consensuado, fue sometido al proceso de revisión de HUPO-PSI para su categorización como estándar. En la definición de este estándar y debido al gran número de casos de uso a incluir en el estándar, *Medina-Aunon* fue el responsable de modelizar y posteriormente validar dos técnicas de proteómica cuantitativa basados en espectrometría de masas: SILAC y ICPL. Igualmente contribuyó junto con el resto de miembros del equipo de trabajo a la definición del esquema y redacción del artículo.

5. PRIDEViewer: a novel user-friendly interface to visualize PRIDE XML files. Medina-Aunon JA, Carazo JM, Albar JP. *Proteomics.* 2011 Jan;11(2):334-7. doi: 10.1002/pmic.201000448.
Índice de impacto 4.132/4.223 (Q1)[a]. Cat. CO. Rank. 14/74.

El repositorio PRIDE rápidamente se convirtió en un referente en cuestión de compartir públicamente los resultados proteómicos. Sin embargo, la revisión de estos ficheros no era un trabajo trivial. *Medina-Aunon*, ideó e implementó en su totalidad la herramienta PRIDEViewer que permitía una revisión completa de la información contenida en un fichero PRIDE XML y la validación de los resultados relativos a identificación de proteínas y péptidos. Para dicho fin, tuvo que diseñar el primer modelo de clases del formato PRIDE XML y además comprender como enlazar y visualizar todas las secciones que contienen. Al margen de la visualización, también diseño una interfaz del motor de búsqueda Mascot. Para ello, tuvo primeramente que idear una interfaz basada en el protocolo TCP (http) que conectase con el servicio de búsqueda y en segundo lugar comprender el envío/recepción de paquetes en relación al proceso de identificación de proteínas y péptidos. Lideró el proceso de escritura y revisión tras los comentarios de los revisores. Actualmente es el encargado de su mantenimiento.

6. Protein Information and Knowledge Extractor: Discovering biological information from proteomics data. Medina-Aunon JA, Paradela A, Macht M, Thiele H, Corthals G, Albar JP. *Proteomics.* 2010 Sep;10(18):3262-71. doi: 10.1002/pmic.201000093.
Índice de impacto 4.132/4.223 (Q1)[a]. Cat. CO. Rank. 14/74

A pesar de que existían servicios de recuperación de anotaciones biológicas en relación a un conjunto de genes, no existía una herramienta similar orientada para proteómica. Tras la lectura y revisión de la mayoría de los servicios gratuitos que se ofrecían para el caso genómico y los casos de uso que se establecieron tras la consulta a investigadores del campo, *Medina-Aunon*, ideó e implementó en su totalidad el servicio PIKE (*Protein Information and Knowledge Extractor*) que permitía reunir información biológica, funcional y patológica relacionada con un conjunto de proteínas. Aunque toda la arquitectura se describe más detalladamente en la sección de resultados, la herramienta resuelve cuestiones típicas relacionadas con la variedad de identificadores de proteínas, su mapeo entre las distintas bases de datos y la definición de flujos de extracción de una forma totalmente dinámica y configurable según la disponibilidad de las bases de datos

que son consultadas. Se incluyen más de una veintena de bases de datos como fuente de datos que fueron estudiadas individualmente para la recuperación del dato de interés y su protocolo de conexión y extracción. Al margen de estas cuestiones, presenta como características el acceso a datos es tiempo real sin el uso de bases de datos intermedias, la presentación integrada de los datos y la posibilidad de obtener distintas vistas de las anotaciones recuperadas.

Además incluyó una interfaz a PRIDE XML para poder complementar un fichero de resultados de este tipo con las anotaciones recuperadas y así ofrecer una visión más completa desde el punto de vista biológico. Lideró el proceso de escritura y revisión del artículo. Actualmente es el encargado de su mantenimiento.

7. In silico analysis of protein neoplastic biomarkers for cervix and uterine cancer. Rodríguez-Pérez MA, **Medina-Aunon A,** Encarnación-Guevara SM, Bernal-Silvia S, Barrera-Saldaña H, Albar-Ramírez JP. Clin Transl Oncol. 2008 Oct;10(10):604-17.
Índice de impacto: 1.276

Durante el tiempo de desarrollo del servicio PIKE, surgió una colaboración en donde se solicitaba a *Medina-Aunon* que estudiase los repositorios y bases de datos públicas con información patológica relacionada con el tema principal del trabajo. Tras la revisión del estado del arte de la tarea encomendaba, *Medina-Aunon,* empleó el sistema *Human Protein Atlas*, para recuperar todos los potenciales biomarcadores para cáncer cérvico-uterino según los distintos estadios y fases definidas en el trabajo y empleó la versión en desarrollo de PIKE para recuperar las anotaciones biológicas y funcionales de interés y así determinar más evidencias que soportasen la candidatura como biomarcadores de las proteínas recopiladas. Para que los datos a estudiar fuesen realmente relevantes, nuevas bases de datos de consulta se estudiaron e incluyeron en el proceso de recuperación de anotaciones. Contribuyó a la escritura del artículo y su revisión.

4.2. Otras publicaciones citadas

8. A guide for integration of proteomic data standards into laboratory workflows. Medina-Aunon JA, Krishna R, Ghali F, Albar JP, Jones AJ. *Proteomics.* 2013 Feb;13(3-4):480-92. doi: 10.1002/pmic.201200268. Epub 2013 Jan 15. *Review.*
Índice de impacto 4.132/4.223 (Q1)[a]. Cat. CO. Rank. 14/74

Tras una relativa trayectoria en el campo de la estandarización y como miembro de HUPO-PSI, *Medina-Aunon* fue invitado por el editor de Proteomics a enviar una revisión durante el año 2012 para el número anual de revisiones de 2013. Se optó por una revisión del estado del arte de las distintas iniciativas, desarrollos y herramientas que permitiese integrar de una manera más sencilla los estándares de datos en los flujos de trabajo de los laboratorios de proteómica. Tras enviar la propuesta de la temática y su aceptación por el editor, recopiló y estudió las distintas opciones que se brindaban para poder llevar a cabo una integración real. Ésta contemplaba el simple uso de herramientas aisladas en pasos sucesivos que dirigen a la obtención de resultados estandarizados según distintos casos de uso a la integración de librerías y APIs (*Application Programming Interface*) en desarrollos propios que los usuarios más expertos puedan crear en el futuro. Parte del trabajo de esta revisión ha sido plasmada en la introducción, concretamente en la sección de estándares de datos proteómicos.

9. Semi-automatic tool to describe, store and compare proteomics experiments based on MIAPE compliant reports. Martínez-Bartolomé S, **Medina-Aunon JA**, Jones AR, Albar JP. *Proteomics.* 2010 Mar;10(6):1256-60. doi: 10.1002/pmic.200900367.
Índice de impacto 4.132/4.223 (Q1)[a]. Cat. CO. Rank. 14/74

Medina-Aunon colaboró en la definición del esquema de la base de datos que almacena la información relativa a los experimentos y su diseño. Colaboró en las pruebas que se realizaron y en la inserción de datos. Contribuyó a la escritura del artículo y se revisión.

10. A Spanish human proteome project: dissection of chromosome 16. Segura V, **Medina-Aunon JA,** Guruceaga E, Gharbi SI, González-Tejedo C, Sánchez del Pino MM, Canals F, Fuentes M, Casal JI, Martínez-Bartolomé S, Elortza F, Mato JM, Arizmendi JM, Abian J, Oliveira E, Gil C, Vivanco F, Blanco F, Albar JP, Corrales FJ. *J Proteome Res.* 2013 Jan 4;12(1):112-22. doi: 10.1021/pr300898u.
Índice de impacto 5.056/5.223 (Q1)[a]. Cat CO. Rank. 10/74

Medina-Aunon definió los flujos de trabajo bioinformáticos y realizó el análisis conjunto de experimentos basados en MS/MS. Contribuyó en la redacción del artículo y posterior revisión.

11. Guidelines for reporting quantitative mass spectrometry based experiments in proteomics. Martínez-Bartolomé S, Deutsch EW, Binz PA, Jones AR, Eisenacher M, Mayer G, Campos A, Canals F, Bech-Serra JJ, Carrascal M, Gay M, Paradela A, Navajas R, Marcilla M, Hernáez ML, Gutiérrez-Blázquez MD, Velarde LF, Aloria K, Beaskoetxea J, **Medina-Aunon JA,** Albar JP. J Proteomics. 2013 Mar 14. doi:pii: S1874-3919(13)00102-4. 10.1016/j.jprot.2013.02.026.
Índice de impacto 4.088/4.302 (Q1)[a]. Cat. CO. Rank. 15/74

Medina-Aunon participó en la coordinación y distribución de trabajo y casos de uso. Participó en la guías que se consensuaron en la reunión de dichos casos de uso y participó en la redacción final de las guías MIAPE.

[a] Datos extraídos de Thomson Reuters 2012. Índice de impacto último año/últimos 5 años. Q: cuartil. Categoría: Biochemical Research Methods. Rank. Posición dentro de la categoría.

* Publicaciones del grupo de trabajo definido por HUPO-PSI. En este tipo de publicaciones los autores se agrupan por nacionalidades, independientemente de su aportación específica, salvo el representante elegido (primer autor) y el responsable del grupo de trabajo (último autor).

Objetivos

Atendiendo al ámbito multidisciplinar del área a cubrir, la elaboración de esta tesis doctoral presenta los siguientes objetivos:

1) Estudio e implantación de estándares internacionales para la representación de datos según las diferentes fases que componen los experimentos proteómicos.

2) Establecer nuevos métodos y herramientas que permitan una visión integral de los experimentos basados en espectrometría de masas y anotados según los estándares internacionales.

3) Establecer nuevos métodos para la extracción de información a partir de bases de datos biológicas y el conjunto de proteínas identificadas en un experimento. En consonancia con los objetivos anteriores, estos métodos serán totalmente compatibles con los resultados obtenidos en 1) y 2).

Materiales y métodos

En la elaboración de los distintos artículos que componen esta tesis doctoral son varios los materiales y métodos que se han empleado. Estos se pueden agrupar en dos grupos atendiendo su naturaleza analítica o bioinformática.

Materiales y métodos analíticos

Selección y preparación de muestras

Distintas muestras provenientes tanto de fuentes elegidas aleatoriamente como extraídas expresamente para la elaboración de los artículos propios de esta tesis doctoral fueron seleccionadas.

Estudio DIGE

En el estudio realizado en la plataforma DIGE, fueron analizadas muestras obtenidas en dos condiciones distintas temporales de exposición, 24h y 96h, en relación a los efectos de un agente anabólico (salbutamol). Las muestras control y tratadas fueron marcadas con los fluoróforos Cy3 y Cy5 siguiendo el protocolo descrito por el fabricante (GE Healthcare, EEUU)[1]. Una mezcla equimolar de todas las muestras fue marcada con

1 Protocolo completo disponible en: http://www.biomedcentral.com/content/supplementary/1756-0500-4-86-s1.doc

Cy2 y empleada como estandar interno. Las muestras fueron depositadas en tiras (*immobiline DryStrips*) de pH 3-10, de rango no linear de 24 centímetros (GE Healthcare, EEUU). La primera dimensión, IEF, fue realizada en Ettan IPGPhor II® (GE Healthcare, EEUU) siguiendo las indicaciones del frabricante. Para la segunda dimensión, las tiras fueron aplicadas en geles de poliacrilamida 12.5% SDS-PAGE según el protocolo descrito en el manual Ettan DIGE® (GE Healhtcare, EEUU).Tras la ejecución de la electroforesis se realizó la correspondiente adquisición de imágenes con el sistema de adquisición Ettan DIGE Imager® (Ge Healthcare, EEUU). Los valores de longitud de onda fijados:

- excitación (Cy3, Cy5 y Cy2): 540/25, 635/30 y 480/30 nm respectivamente.
- emisión (Cy3, Cy5 y Cy2): 595/25, 680/30 y 530/30nm respectivamente.

Finalizada la fase de adquisición, se llevó a cabo el análisis de las imágenes resultantes con el software de análisis diferencias DeCyder® (GE HealthCare, EEUU). De las 12 imágenes adquiridas (3 canales (Cy3, Cy5 y Cy2) x 4 geles) se seleccionaron como spots de interés aquellos que se visualizabna en 9 de las 12 imágenes y con un índice de variación (*fold change*) igual a superior a 1.5.

Siguiendo este protocolo y en base a las dos condiciones temporales establecidas (24 y 96h.) se aislaron dos conjuntos de spots de 31 y 110 respectivamente. Estos fueron recortados, digeridos con tripsina y analizadas a posteriori por espectrometría de masas MS/MS.

Muestras ABRF

Para parte de los experimentos que sirvieron de base para la realización de los trabajos que forman parte de esta tesis doctoral se seleccionaron las muestras de la participación del laboratorio de Proteómica del Centro Nacional de Biotecnología (Madrid, España) en el estudio internacional del año 2010 promovido por la asociación de servicios de recursos biomolecualres (ABRF: *Association of Biomolecular Resource Facilities*). El objetivo y la composición de la muestra de este estudio se recoge en el sitio web del grupo de investigación de estandards en Proteómica[2].

Muestras servicio rutinario

Por último también se seleccionaron muestras aleatorias provenientes de estudios reales realizados en el laboratorio de Proteómica del Centro Nacional de Biotecnología (Madrid, España) de diversas especies y condiciones. Para preservar la privacidad del origen de etas muestras y no poder vincularlas con los resultados reunidos, se ha optado por no indicar más información al respecto de estas muestras.

Espectrometría de masas

Espectrometría de masas MALDI-TOF

Las muestras analizadas por la plataforma de espectrometría de masas MALDI TOF/TOF (MS+MS/MS) fueron digeridas en gel por tripsina en el digestor automático Proteineer dp® (Bruker Daltonics, Bremen, Alemania). Una vez digeridas, los distintos complejos de péptidos trípticos fueron depositados en la placa

2 http://www.abrf.org/index.cfm/group.show/proteomicsstandardsresearchgroup.47.htm.

porta muestras y enlazados con la matriz alfa-ciano-4-hidroxi-cinámico. La obtención de la huella peptídica de distintas proteínas empleadas se obtuvieron con el espectrómetros de masas MALDI TOF/TOF AB 4800 (ABSciex, MA, EEUU), y analizados con el software AB Data Analyst v3.2 (ABSciex,MA, EEUU). Los distintos ficheros que contenian las listas de picos fueron exportados al formato por defecto de Mascot, (-mgf: *Mascot General File)* para su posterior identificación.

Espectrometría de masas Tándem (LC-MS/MS)

La muestra del estudio ABRF2010, fue analizada utilizando el cromatógrafo 3000 nano-HPLC-Dionex (Thermo Fisher Scientific, CA, EEUU) acoplado al espectrómetro de masas tanden, HCT Ultra Ion Trap (Bruker Daltonics, Bremen, Alemania). El análisis, basado en un experimento dependiente de datos (*data-dependent*) realizaba una selección de los iones parentales de mayor abundancia en modo MS, para posteriormente realizar su fragmentación, alternando los modos de disociación inducida por colisión (CID: *collision-induced dissociation*) y disociación por transferencia de electrones (ETD: *electron transfer dissociation*).

Por su parte para la muestra que sirvió de base al experimento DIGE se empleó un cromotrógrafo idéntico al caso anterior que se acoplaba a un un espectrómetro de masas LTQ (Thermo Fisher Scientific, CA, EEUU). En este caso los peptídos ionizados fueron seleccionados según el modo dependiente de datos denominado "*triple-play*", en donde se seleccionaban los tres iones de mayor abundancia para su posterior fragmentación en modo inducida por colisión.

Materiales y métodos bioinformáticos

Motores de búsqueda

Los motores de búsqueda y los parámetros empleados en la fase de identificación de péptidos y proteínas en la elaboración de esta tesis doctoral son:

Motor de búsqueda: Mascot y Phenyx

En la fase de identificación de péptidos y proteínas, todos los datos de espectrometría de masas obtenidos fueron analizados por el software Mascot (MatrixScience, Londres, Reino Unido). Diferentes versiones, 2,1, 2.2 y 2.3, de esta herramienta fueron utilizadas tanto para la identificación de proteínas por huella peptídica (PMF) como por la identificación de péptidos por espectro de fragmentación (PFF). En una serie de experimentos incluidos en las publicaciones, los espectros MS/MS fueron analizados con el motor de búsqueda Phenyx (GeneBio, Ginebra, Suiza) para complementar los resultados obtenidos por Mascot. La versión del software utilizada fue la 2.5. La tabla M1.1 contiene un resumen de los parámetros empleados en la búsqueda según el experimento.

CONFIGURACIÓN/PARAMÉTRO	DIGE	ABRF	STD. (MS +MS/MS)
Base de Datos	IPI v 3.53	UniProt/SwissProt 2010_12	UniProt/SwissProt (*Forward+Decoy*)*
Taxonomia	Roedores (Rodetia)	Mamíferos (Mammalia)	Variable **
Enzima	Tripsina	Tripsina	Tripsina
Errores de corte (misscleavage)	1	1	2
Tolernancia Péptido	+/- 1.5 Da	+/- 0.5 Da	+/- 50 ppm
Tolerancia MS/MS	+/- 0.5 Da	+/- 0.6 Da	+/- 50 ppm o [0.2-0.6] Da
Modificaciones Fijas	CAM (Cys)	--	CAM (Cys)
Modificaciones variables	Ox (Met)	Ox (Met) Phos (Ser, Tyr, Thr)	Ox (Met) Phos (Ser, Tyr, Thr)
Medición	Monoisotópica	Monoisotópica	Monoisotópica

*: Diferentes versiones.
**: Habitualmente humana (Homo sapiens).

Tabla M1.1: Resumen de parámetros de búsqueda agrupados por origen de la muestra.

Cabe destacar de este proceso que en ciertas ocasiones y con el objetivo de poder estimar la tasa de error que se desprende de la identificación se emplean bases de datos señuelos o *Decoy*. Simplificando, estas bases de datos combinan entradas reales de proteínas (*forward*) y otras aleatorias o falsas (*decoy*) que se marcan con entradas señuelos. Por la naturaleza del proceso de identificación, habrá unas pocas ocasiones en donde un espectro se empareje con una entrada señuelo. Una vez que se tiene el listado total de proteínas identificadas, se puede analizar la salida y comprobar en dicho listado cuantas entradas falsas aparecen. Esto proporciona una medida en forma de ratio de entradas verdaderas versus entradas falsas, que a su vez sirve para evaluar la confianza del resultado final.

Estándares de datos

El uso de estándares de datos, normalmente formateados como documentos XML, y sus correspondientes guías MIAPE, es un elemento troncal para eliminar, en la medida de lo posible, la dependencia de la plataforma analítica

MzML y MIAPE MS

Los datos resultantes del análisis por espectrometría de masas fueron exportados al estándar propuesto por HUPO PSI mzML (versión 1.0 y versión 1.1) (Martens, Chambers et al. 2011). Adicionalmente, y en aquellos casos que resultó necesario, se generaron los correspondientes informes normalizados MIAPE-MS (Taylor, Binz, et al. 2008), en donde se incluían los diferentes metadatos relativos a al adquisición.

MzIdentML, PRIDE XML y MIAPE MSI

Por su parte, los datos relativos a la identificación de péptidos y proteínas, fueron exportados al estándar propuesto por HUPO PSI mzIdentML (versiones 1.0 y 1.1) (Jones, Eisenacher et al. 2012) y sus correspondientes documentos MIAPE-MSI (Binz, Barkovich et al. 2008). Por último y para poder enviar los resultados a un repositorio externo para su evaluación, fueron depositados en el repositorio PRIDE (http://www.ebi.ac.uk/pride) según el formato propuesto PRIDE XML (versión 2.1).

BLOQUE 2: RESULTADOS, DISCUSIÓN Y CONCLUSIONES

R1: Estudio e implantación de estándares internacionales para la representación de datos según las diferentes fases que componen los experimentos proteómicos

En relación con este objetivo son varios los trabajos incorporados en esta tesis doctoral que muestran una sustancial mejora en el campo de aplicación. Sin embargo, a la hora de clasificarlos, son dos las opciones que pueden considerarse. La primera de estas se debe a su propósito, agrupándose en dos bloques:

a) Aquellos que presentan un nuevo esquema o formato para el reporte de los datos generados según alguna de las fases en que se dividen los experimentos proteómicos.
b) Aquellos que presentan un marco de trabajo donde poder importar, exportar y manipular los datos derivados de los experimentos según los formatos estándar.

La segunda clasificación atiende al área de aplicación en donde se pueden diferencias las siguientes tres categorías:

a) Separación de proteínas por electroforesis.
b) Espectrometría de masas e identificación de péptidos y proteínas.
c) Proteómica diferencial cuantitativa.

En un intento de combinar ambos criterios y con el fin de presentar los avances de la forma más similar a la metodología de trabajo experimental en proteómica, a continuación, y en este orden, se resumen los trabajos relativos a:

1) Formato estándar para experimentos basados en separación por electroforesis (GelML).
2) Nuevo entorno de trabajo para el reporte de experimentos basados en electroforesis bidimensional utilizando el formato PRIDE XML.
3) Entorno de trabajo *ProteoRed MIAPE Web Toolkit* (PMWTK) para la integración de los estándares relativos a las fases de electroforesis (GelML), espectrometría de masas (mzML) e identificación de péptidos y proteínas (mzIdentML y PRIDE XML) en las rutinas de trabajo del laboratorio.
4) Formato estándar para cuantificación de péptidos y proteínas identificados por espectrometría de masas: mzQuantML.

R1.1. Formato de intercambio de experimentos basados en electroforesis: GelML

El formato GelML (Gibson, Hoogland et al. 2010) ha sido fruto de una colaboración entre varios integrantes del grupo de trabajo HUPO-PSI. GelML modela el proceso de separación electroforética de proteínas y su aplicación en el contexto de un experimento proteómico tiene lugar una vez finalizada la preparación de la muestra y previo al análisis de imagen y la identificación de proteínas por espectrometría de masas.

Este modelo soporta la descripción de los protocolos que se aplican como son la ejecución de la electroforesis, la detección de proteínas, tanto directa como indirectamente y la adquisición de imágenes del gel resultante. Aunque no incorpora secciones dedicadas a la preparación o procesado de la muestra, dicha información si se puede capturar con el modelo FuGE (*Functional Genomics Experiment* (Jones, Miller et al. 2007), el cual se integra dentro del esquema GelML.

Aunque este esquema no proporciona un soporte detallado del proceso del análisis de imagen, si permite incorporar la localización de los spots identificados en las imágenes resultantes así como una mínima información relativa a la cuantificación de dichos spots.

Modelo GelML

A continuación se presentan las secciones en las cuales se divide el modelo GelML. Cada una representa un paso diferenciado de la separación por electroforesis de una experimento:

- Composición del gel y descripción del mismo en el caso que se traten de geles prefabricados.
- Protocolos de electroforesis que puede ser monodimensional (1-DE), bidimensional (2-DE) o procesos no estándar como la electroforesis tridimensional (3-DE) (Rabilloud 2013).
- Carga de la muestra.
- Ejecución de la electroforesis.
- Tinción y detección de proteínas.

- Adquisición de imágenes .
- Corte de bandas (1-DE) o spots (2-DE).

Como anteriormente se indicaba, GelML extiende elementos del modelo FuGE. Concretamente, los elementos relativos a:

- Los protocolos o procedimientos aplicados (*Protocol)*.
- La ejecución de estos protocolos y los parámetros de ejecución (*ProtocolApplication*).
- Todos los materiales físicos y biológicos (*Material*).
- Los ficheros de datos (*Data*).

Los ficheros GelML resultantes para la aplicación clásica de la electroforesis tanto en 1D como en 2D son muy similares, diferenciándose únicamente en el elemento reservado para la definición de dimensiones, y como es obvio, la ejecución de la segunda dimensión. Un esquema de la aplicación de la electroforesis bidimensional (2-DE) se presenta en la figura R1.1.

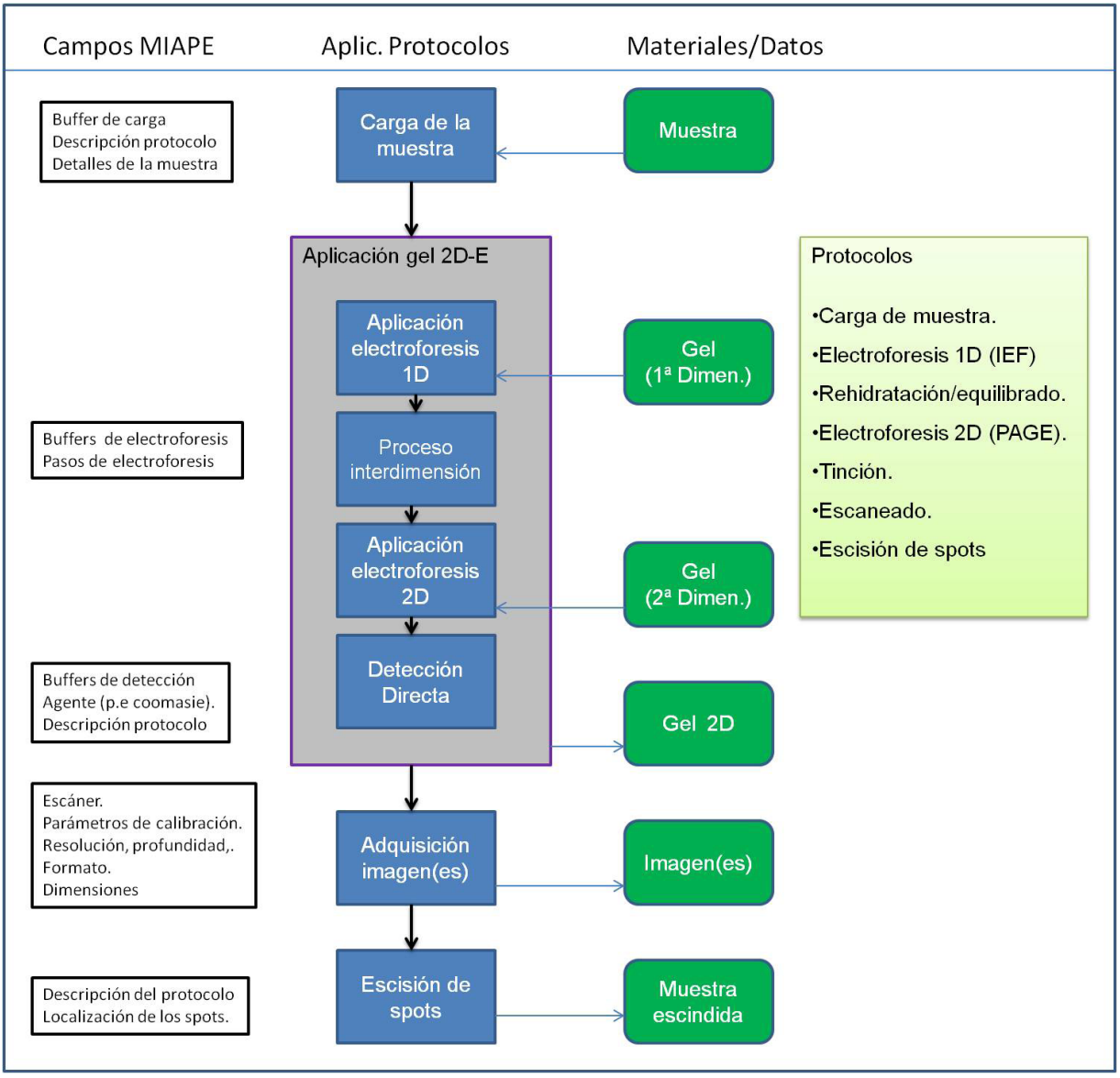

Figura R1.1: Esquema GelML. Elementos que describen el proceso de separación de prtoteínas 2D-E. Adaptación de (Gibson, Hoogland et al. 2010).

El eje central del fichero de resultados es una serie de ejecuciones de los protocolos (elementos *ProtocolApplications* representados por rectángulos en la figura R1.1) y cuya función es enlazar las entradas y salidas de los distintos pasos a realizar. Tanto las entradas como las salidas para cada elemento *ProtocolApplication* sólo pueden ser de dos tipos: Materiales (elemento *Material*) tanto físicos como biológicos y datos (elemento *Data*) que podemos interpretar como resultados, tanto intermedios como finales. Ambos elementos se representan como rectángulos con esquinas redondeadas. Esta estructura permite cierta flexibilidad en la forma de aplicar procedimientos no estándar dentro de los flujos de trabajo.

Cada elemento *ProtocolApplication* debe referenciar al correspondiente Protocolo (elemento *Protocol*) que se define en el mismo fichero. Cada elemento protocolo incluye una descripción del mismo, redactado en lenguaje natural, el equipamiento y/o software y parámetros empleados en su ejecución. Esta organización permite ahorrar en el número de definiciones de los protocolos, ya que aunque se aplique en varias ocasiones (en un elemento *ProtocolApplication*), únicamente habrá que definirlo en una ocasión.

Adicionalmente, la figura R1.1 muestra como diferentes elementos clave para las guías MIAPE (MIAPE-GE (Gibson, Anderson et al. 2008)) son capturados en cada etapa. Tres ficheros completos ejemplificando los tres escenarios más frecuentes, 1-DE, 2-DE y DIGE, de la separación de proteínas por electroforesis se presentan como material suplementario.

Implementación GelML

Como parte de la aportación de este trabajo se desarrolló el primer sistema que integraba GelML y una base de datos relacional. A partir de los formularios que se definieron para entrada de datos en la base de datos de documentos MIAPE (ProteoRed MIAPE Database: PMDB) (Martinez-Bartolome, Medina-Aunon et al. 2010), cualquier usuario puede introducir la información más relevante, y a su vez la exigida por las guías MIAPE, relativa a la ejecución de la fase de electroforesis. La herramienta desarrolladada en Java (www.java.com) permite exportar los datos almacenados en la BD en el formato GelML. A través de unos sencillos pasos que indica el asistente de la aplicación, los datos relativos a una separación electroforética serán primeramente seleccionados, posteriormente validados y finalmente exportados en un fichero de resultados siguiendo el nuevo esquema estándar GelML.

R1.2. Nuevo entorno de trabajo para el reporte de experimentos basados en electroforesis bidimensional utilizando el formato PRIDE XML

La generación de ficheros PRIDE XML no es una cuestión trivial para experimentos proteómicos, incluso aquellos que sólo contienen sólo unas decenas de proteínas. Por esta razón se requiere software, que partiendo de ficheros generados automáticamente en las distintas fases del experimento, permita interpretarlos, combinarlos y finalmente generar el resultado según el formato PRIDE XML.

Una de estas herramientas es el PRIDE Converter (Barsnes, Vizcaino et al. 2009). Este software permite la conversión de distintos formatos de ficheros que incluyen los espectros de masas e identificación de péptidos y proteínas provenientes de los motores de búsqueda en el formato PRIDE XML y su posterior envío a la base de datos PRIDE.

PRIDE Converter y datos de electroforesis

A pesar de resultar una herramienta muy útil para el reporte de proteínas en formato estándar, PRIDE Converter es una herramienta que se ha diseñado para experimentos basados en separación de péptidos sin hacer uso de electroforesis. Esta aproximación, no realiza una separación previa de proteínas (como si hace la electroforesis) y analiza todos los péptidos al mismo nivel. Este hecho requiere la aplicación posterior de la denominada inferencia de proteínas (*protein inference*) (Huang, Wang et al. 2012; Prieto, Aloria et al. 2012) que consiste en determinar la relación entre los péptidos realmente identificados y las proteínas a las que supuestamente pertenecen dichos péptidos. Este enfoque no se ajusta a los estudios basados en electroforesis, debido a que en estos últimos, la proteína a analizar siempre se aísla, y por tanto los péptidos identificados siempre mantienen esta relación presente.

Más en detalle, PRIDE Converter, es capaz de procesar dentro de un mismo experimento varios ficheros resultantes de los motores de búsqueda. En el caso de electroforesis estos resultados presentarían una asociación uno a uno con un spot y por extensión con su hipotética proteína de origen. Sin embargo, la lógica interna de este software obvia esta relación y reduce el conjunto de proteínas resultantes a las que se infieren a partir de una lista combinada de todos los péptidos identificados.

En el caso de estudios basados en gel, el aislamiento de la proteína relaciona directamente un spot dentro de la matriz con el fichero resultante del motor de búsqueda, asegurando que los péptidos identificados pertenecen a esa proteína asilada, y por extensión a ese spot, y no a otra. Este hecho evita la inferencia de proteínas debido a los péptidos que provienen de un determinado spot no deben combinarse con los correspondientes a otros spots. Además, PRIDE Converter no proporciona ningún método para subir imágenes del gel obtenido y analizado, incorporar coordenadas de los spots detectados y posteriormente identificados o información adicional relativa a la cuantificación de proteínas basada en gel.

PRIDEConverter y PRIDESpotMapper

Las limitaciones descritas y otras menos significativas derivó en una colaboración con el equipo responsable de todo lo relativo a PRIDE en el Instituto Europeo de Bioinformática (*European Bioinformatics Institute*, EBI). En concreto, los integrantes de este equipo proporcionarían una versión adaptada del software PRIDE Converter en donde cada péptido identificado se anotaría con el nombre del fichero de identificación al que pertenece. Este dato adicional se utilizaría para establecer la relación de un spot por un fichero de identificación.

La otra parte de la colaboración consistió en la creación de una nueva aplicación, llamada PRIDESpotMapper (Medina-Aunon, Kenyani et al. 2011), que complementase la salida de la versión adaptada de PRIDE Converter y permitiese introducir la información relativa a cada uno de los spots identificados en un determinado gel, como por ejemplo imagen del gel o coordenadas de localización dentro de la imagen.

La lógica de esta nueva aplicación desarrollada en Java consiste en modificar el fichero PRIDE XML generado por la versión adaptada de PRIDE Converter y aislar cada uno de los péptidos identificados. Posteriormente, estos péptidos se agrupan en dos niveles, el primero por la proteína a la que se asocian y el segundo según el fichero de resultados al que dicho péptido pertenece. Este doble agrupamiento permite poder separar proteínas que comparten un conjunto de péptidos, pero adicionalmente y gracias a la información que proporciona la separación por electroforesis, los péptidos que pertenece a cada spot aislado. La información necesaria para establecer la relación entre los ficheros de resultados, uno por spot, y la información derivada de la separación por electroforesis para cada uno de esos spots se lleva a cabo por una hoja de cálculo (por ejemplo de Microsoft Excel©) que se puede generar fácil y prácticamente de una manera automática a partir de los formatos de salida proporcionados por las principales aplicaciones de análisis de imagen.

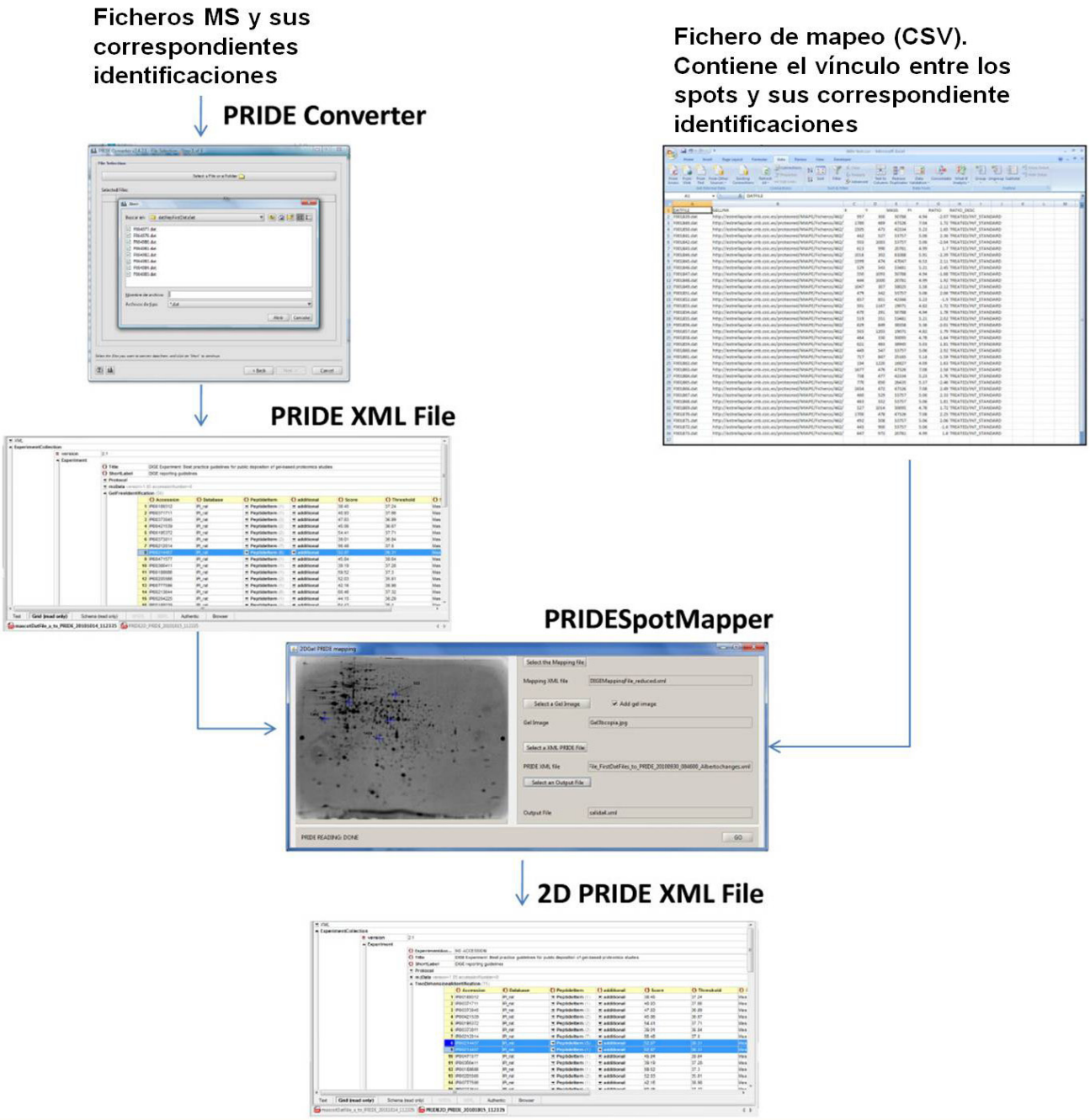

Figura R1.2: Esquema de trabajo de PRIDESpotMapper. Inclusión de datos 2D-E en el formato PRIDE XML. Adaptación de (Medina-Aunon, Kenyani et al. 2011).

El flujo de información (Fig. R1.2) es directo y se realiza en unos pocos pasos. Primero se selecciona y carga la hoja de cálculo que contiene el mapeo entre cada fichero de resultados del motor de búsqueda y la información del spot al que se asocia (relación resultado - spot). Segundo, opcionalmente se puede indicar la URL o la ruta local de la imagen del gel, simplemente para comprobar la ubicación de los spots descritos en la hoja de cálculo. Como tercer y último paso se selecciona el fichero PRIDE XML generado a partir de la reunión de todos los ficheros de resultados tras la ejecución adaptada de PRIDE Converter. Con estos parámetros de entada, la aplicación fusionará los ficheros de datos para generar un nuevo fichero PRIDE XML, denominado 2D PRIDE XML (material suplementario), en donde cada spot se enlaza con una única proteína con sus correspondientes péptidos, incluyendo las coordenadas de localización dentro del gel y, en el caso de haberla, la información relativa a la cuantificación.

R1.3. Entorno de trabajo ProteoRed MIAPE Web Toolkit para la integración de los estándares relativos a las fases de electroforesis (GelML), espectrometría de masas (mzML) e identificación de péptidos y proteínas (mzIdentML y PRIDE) en las rutinas de trabajo del laboratorio

Una de las principales desventajas a la hora de trabajar con datos provenientes de ensayos proteómicos es la heterogeneidad de las fuentes de datos según las distintas fases o etapas que componene el resultado de un experimento (Fig. R1.3). En este sentido y siguiendo un enfoque simplista, podemos clasificar las fuentes de información según las siguientes tres categorías:

Manual: este tipo de información proviene de las anotaciones que el personal añade en sus cuadernos de laboratorio. Toma de tiempos, concentración o definición de protocolos son algunos de estos ejemplos y presentan la característica que o bien no son procesados por instrumentos analíticos o computacionales, o si lo están, sólo consiste en texto almacenado en lenguaje natural lo que dificulta su aplicación en un entorno computacional. El ejemplo de este tipo de información en un contexto proteómico sería los datos derivados de la electroforesis bidimensional.

Instrumental: Esta categoría engloba aquellos datos que son generados por equipos o instrumentos analíticos, requiriendo el uso de computadores únicamente para la traducción de señales y el almacenaje de datos. Ejemplo de este tipo de información sería la proveniente de los espectrómetros de masas.

Computacional: En este caso los datos son introducidos, traducidos y procesados por recursos computacionales. Esta categoría englobaría los paquetes software dedicados a la identificación y/o cuantificación de péptidos y proteínas.

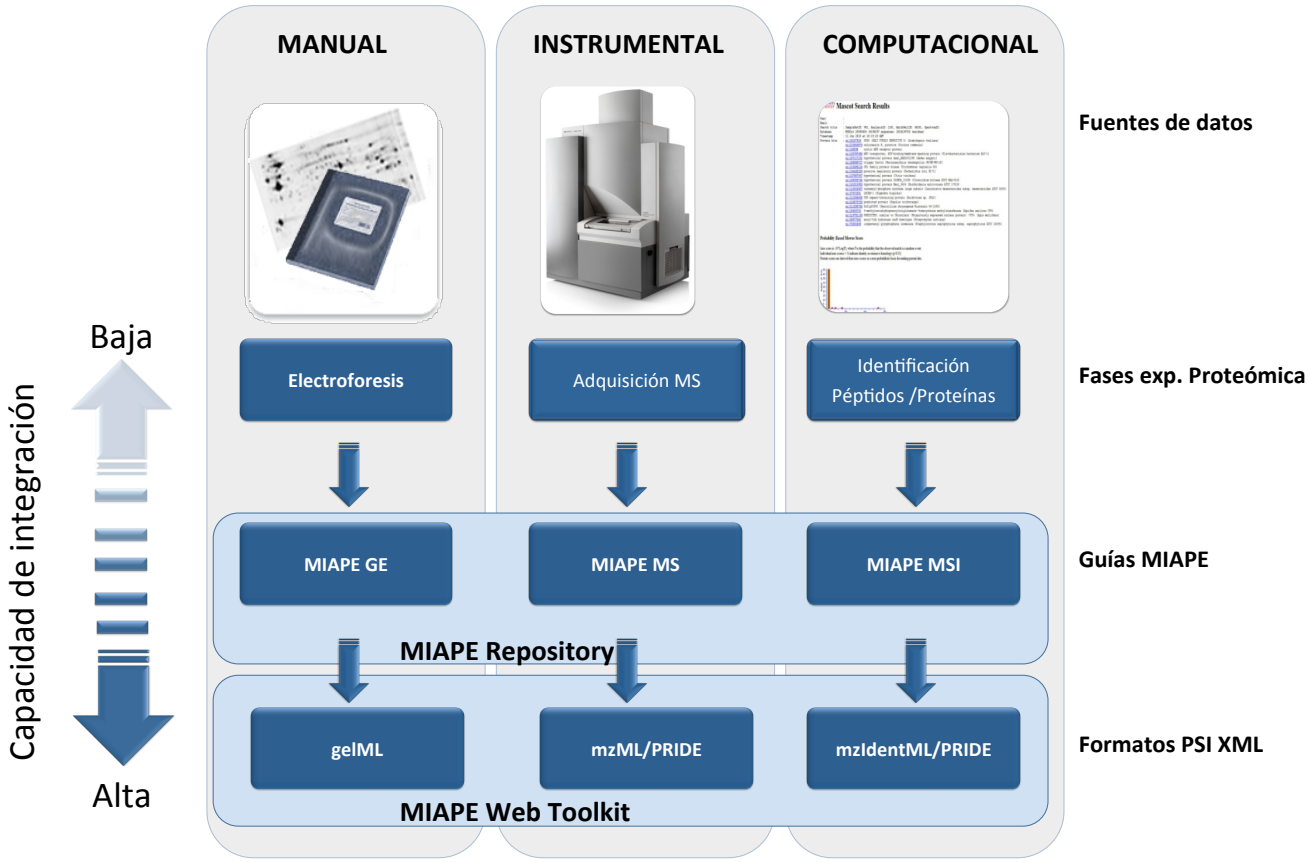

Figura R1.3: Clasificación de los tipos de información que se originan en las fases en que se puede dividir un experimento proteómico. Identificación de cada tipo con su correspondiente guía MIAPE y su formato PSI XML de intercambio. Adaptación de (Medina-Aunon, Martinez-Bartolome et al. 2011).

El entorno de trabajo ProteoRed MIAPE Web Toolkit (PMWTK)

El entorno de trabajo PMWTK (Medina-Aunon, Martinez-Bartolome et al. 2011) permite trabajar con distintas fuentes de datos para, en primer lugar, traducir los distintos datos que se obtienen en las distintas fases de un experimento proteómico a sus respectivos estándares de datos definidos por el HUPO-PSI y sus respectivas guías MIAPE, y por otra, integrar armonizadamente estas fases diferenciables de un experimento en un único resultado siguiendo el formato PRIDE XML. Actualmente PMWTK permite trabajar con las tres fuentes de datos anteriormente descritas y en cuestión de fases, abarca la información generada en la electroforesis, espectrometría de masas e identificación de péptidos y proteínas, restando únicamente el estándar definido para la cuantificación (mzQuantML (Walzer, Qi et al. 2013)) y su correspondiente documento MIAPE (MIAPE Quant (Martinez-Bartolome, Deutsch et al. 2013)).

Para establecer una correspondencia bidireccional entre los principales ficheros de resultados que es capar de manipular el PMWTK (PSI XML, MIAPE y PRIDE XML) se han creado y validado, manualmente en algunos casos, una serie de documentos que mapean los términos y metadatos incluidos en cada formato.

Estructura del PMWTK

PMWTK se apoya en la base de datos de documentos MIAPE (PMDB) (Martinez-Bartolome, Medina-Aunon et al. 2010). Partiendo de la estructura de almacenamiento que proporciona esta base de datos, se desarrollan los tres módulos que componen la estructura del PMWTK (Fig. R1.4).

MIAPE Compliance Checker o validador de contenidos

Uno de los elementos más importantes a la hora de compartir y reportar los datos relativos a un experimento son los metadatos que describen detalladamente la generación de los datos hasta el punto de permitir la reproducción de los mismos. El validador de contenidos permite trasladar estos metadatos entre los documentos MIAPE, descritos usando el lenguaje natural, y a la sintaxis requerida en cada uno de los formatos PSI XML. Para asegurar esta traducción, el validador es invocado tanto a la hora de importar documentos externos en la base de datos MIAPE como a la hora de exportar los registros requeridos de la base de datos para generar los correspondientes ficheros de salida XML (PSI XML y PRIDE XML). Más en detalle, la tarea de validación realizada por este módulo se divide en tres pasos que se ejecutan de forma secuencial y automática:

Primero, se lleva a cabo una validación de contenidos según el contexto del tipo o fase del experimento. Un ejemplo de esta validación para el caso de la separación por electroforesis bidimensional (2-DE), es la comprobación que los datos asociados a los protocolos y sus correspondientes aplicaciones estén definidos para ambas dimensiones.

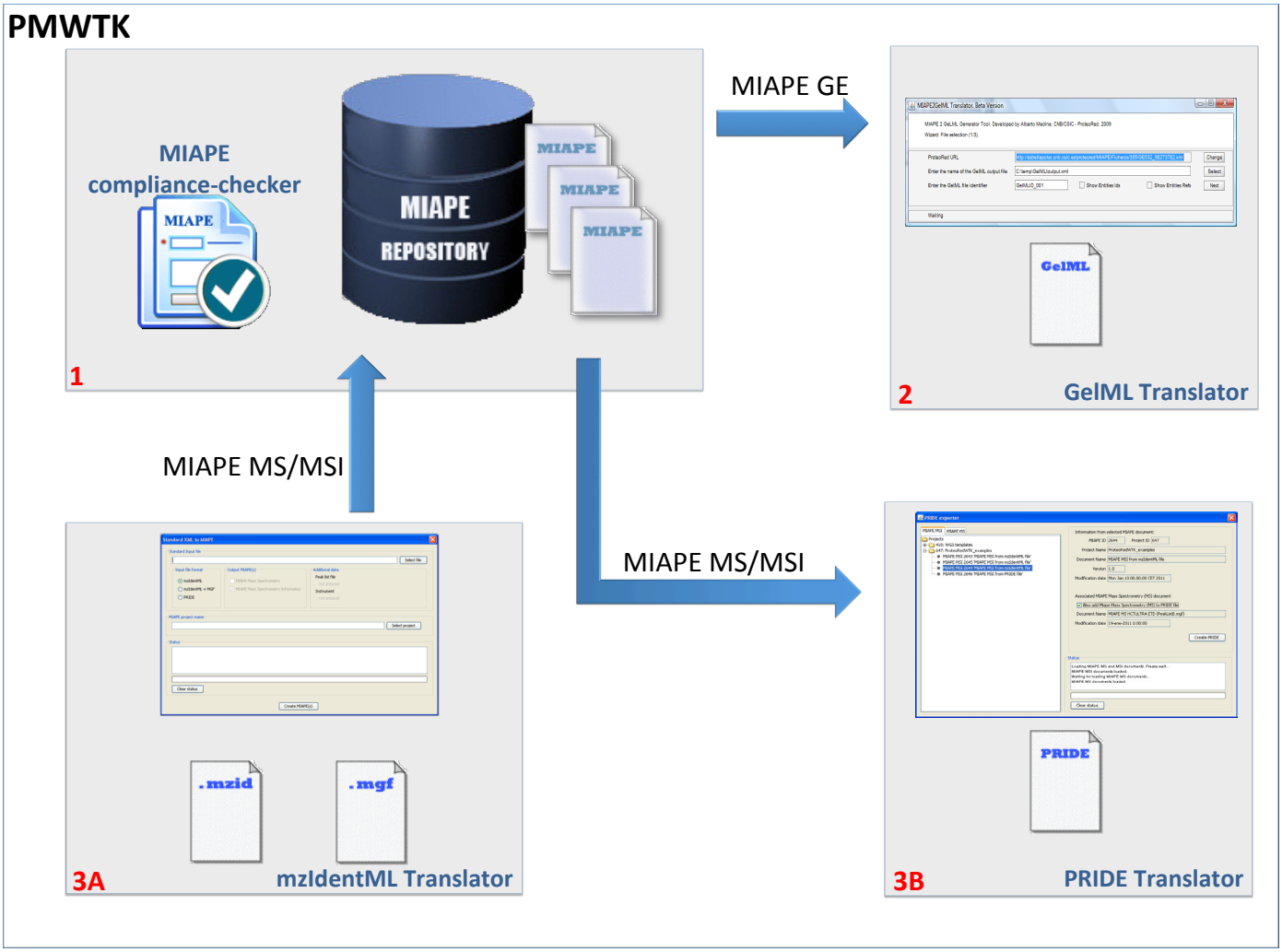

Figura R1.4: Estructura del entorno de trabajo ProteoRed MIAPE Web ToolKit (PMWTK). 1. Base de datos MIAPE (PMDB) y validador de contenidos. 2. Módulo para el tratamiento de datos de separación por electroforesi (GelML Translator). 3A. Traducción estándares PSI XML a guías MIAPE. 3B. Traducción de guías MIAPE a PRIDE XML. Adaptación de (Medina-Aunon, Martinez-Bartolome et al. 2011).

Segundo, se realiza una validación semántica. En este caso, una vez comprobados que los campos necesarios contienen la información requerida, se comprueba que la información realmente describe el campo requerido

a partir de la definición de ontologías y vocabularios controlados. En este caso, cada término del vocabulario controlado es evaluado a partir de las reglas contenidas en el *PSI validator framework* (Montecchi-Palazzi, Kerrien et al. 2009).

En el *tercer* y último paso se comprueban las correspondencias entre elementos incluidos en las distintas partes del documento. En este caso, se comprueban vínculos entre protocolos y la aplicación de estos, como por ejemplo la descripción de la adquisición de una imagen con sus correspondientes parámetros y su imagen resultante.

Datos separación por electroforesis: GelML Translator

Este módulo es el encargado de exportar los datos correspondientes a una separación por electroforesis según el formato GelML desde la base de datos PMDB. La lógica de la aplicación, descrita en los siguientes tres pasos, es directa y bastante simple:

Primero, la aplicación localiza toda la información relevante registrada en un documento MIAPE GE y extrae todos los elementos que intervienen en la descripción del experimento como puede se la matriz del gel, la composición de cada uno de los distintos búferes utilizados o las imágenes de los geles.

Segundo, este módulo genera a partir de elementos adicionales del documento MIAPE aquellas secciones del GelML que no han sido completadas. Cada uno de estos elementos son editables por el usuario por si éste desea detallar más la información.

Tercero, se formalizan las relaciones entre los elementos que forman el recién generado GelML, de tal forma que todo el proceso quede bien integrado y validado. En este paso se comprueban reglas de tipo que la muestra que se separa por la electroforesis tenga tanto su protocolo como su buffer de aplicación y que el resultado de este paso, la matriz del gel, también esté descrita. Esta transformación resulta esencial dentro del documento GelML para asegurar las restricciones de existencia e integridad entre los elementos descritos.

Importar/Exportar datos MS y MSI.

A diferencia de los datos derivados de la fase de separación por electroforesis, los datos generados en las fases de espectrometría de masas (tipo instrumental) e identificación de péptidos y proteínas (computacional) pueden ser tratados computacionalmente de una forma más o menos directa. La base de datos PMDB minimiza el esfuerzo en traducir los ficheros estándar mzML y mzIdentML en los correspondientes documentos MIAPE MS y MSI y su posterior almacenamiento. Esta última característica permite editar o completar cualquier metadato que no estuviese documentado o suficientemente descrito en los ficheros XML iniciales. De forma análoga, a partir de un par de documentos MIAPE MS y MIAPE MSI que pertenezcan al mismo experimento, este módulo se encarga de combinarlos y generar un fichero de resultados PRIDE XML.

R1.4. Formato estándar para cuantificación de péptidos y proteínas identificados por espectrometría de masas: mzQuantML

Las herramientas software tienden a presentar los resultados cuantitativos como un conjunto de tablas o matrices cuyas celdas contienen valores numéricos que representan la abundancia de los péptidos y/o proteínas analizadas. Sin embargo, este tipo de representación no almacena los metadatos que, entre otros, permiten mantener la trazabilidad de cómo la abundancia de esos péptidos y proteínas se ha calculado a partir de los datos de espectrometrías de masas.

Junto con esta característica hay que añadir el bajo control de calidad al que han sido sometidos estas herramientas de cuantificación, originando cierta desconfianza de cómo dichos valores han sido calculados. Como anteriormente se ha descrito, los dos estándares propuestos por el HUPO PSI para espectrometría de masas e identificación de péptidos y proteínas, mzML y mzIdentML, están más o menos integrados en el contexto proteómico, tanto para el intercambio de datos como incluidos en los planes de desarrollo de nuevas herramientas. Sin embargo, no existía un formato estándar que capturase tanto los metadatos como los datos relativos a la fase de cuantificación. Por este motivo, y en el seno del grupo de trabajo correspondiente de HUPO-PSI, se ha colaborado en la definición del estándar mzQuantML (Walzer, Qi et al. 2013).

El formato mzQuantML

El estándar mzQuantML se ha diseñado para permitir una perfecta trazabilidad entre los valores relativos a la variación de abundancia entre los péptidos y proteínas que están presentes en un conjunto de muestras y como esos valores se han calculado a partir de los datos de espectrometría de masas. Este estándar permite la anotación de estos valores a distintos niveles, incluyendo los denominados grupos de proteínas (donde se pretende resolver la ambigüedad que aparece en la inferencia de proteínas) y la relación de estos con las distintas réplicas, condiciones o muestras sometidas a estudio. Estas últimas, incorporadas por primera vez en un formato de datos, muestran la relación entre la fase de análisis y la condición biológica de la que se extraen las muestras analizadas, permitiendo un seguimiento completo del experimento considerando ambos puntos de vista.

Los principios generales que el formato soporta son:

- Conexión con el documento MIAPE Quant.
- Envío de datos cuantitativos a bases de datos públicas especializadas.
- Intercambio de datos cuantitativos entre herramientas software.
- Permitir el análisis estadístico de los datos y la capacidad de reprocesar o recrear el flujo de trabajo utilizando los mismo parámetros experimentales.

Para cumplir dichos principios, el esquema debe capturar los siguientes elementos:

1) Valores finales de abundancia (relativos o absolutos) para péptidos, proteínas y grupos de proteínas.
2) Valores cuantitativos de modificaciones de péptidos y proteínas.
3) Valores de abundancia a nivel de análisis (carrera o *run*) o conjuntos de estos.

4) Pruebas o evidencias que permitan trazar como los valores finales de abundancia fueron calculados.
5) Las relaciones que se establecen entre los elementos utilizados para cuantificar los péptidos y proteínas (*features).*
6) Las relaciones que se establecen entre carreras o *runs* en donde se puede localizar el mismo péptido.
7) Detalles sobre el prefraccionamiento del experimento cuando éste así lo requiera debido a su complejidad.

Para la verificación sintáctica y semántica de los ficheros mzQuantML, y atendiendo a la complejidad asociada a la heterogeneidad de los diferentes métodos cuantitativos, junto con la definición del esquema del formato (XSD) y las incorporaciones de nuevos términos y definiciones a la ontología PSI-MS[1] se han definido una serie de reglas semánticas que permitan diferenciar los distintos tipos de cuantificación: basados en intensidad sin marcaje (*label-free*), marcaje MS^1, marcaje MS^2 y contabilización de espectros (*spectral counting*). Estas reglas se incorporarán a aquellas herramientas software que quieran incorporar este formato como salida estándar de la cuantificación.

El modelo mzQuantML

El modelo captura los metadatos de cómo se ha llevado a cado el análisis cuantitativo y la descripción del diseño experimental, incluyendo secciones dedicadas a las réplicas, técnicas y biológicas, y el agrupamiento de estas en lo que se ha definido como las variables de estudio.

Estos aspectos son importantes de capturar debido a la falta de conocimiento sobre la realidad biológica analizada, consecuencia del modo en que algunos paquetes software tratan esta información. En muchos casos esta se reduce a un resumen efectuado mediante una simple media aritmética a través de todas las réplicas que se asocian a una condición biológica del experimento. El formato define matrices para capturar los valores a varios niveles, desde los *features* individuales hasta las proteínas o grupos de proteínas cuantificados por el software.

Metadatos y parámetros del software

Como se muestra en la figura R1.5, el fichero captura los metadatos a partir de la definición de los vocabularios controlados (CVs) incluidos en la sección< *CvList*>, el creador del fichero <*Provide*r> y los datos necesarios para contactar con este último <*AuditCollection*>. Un fichero considerado como válido debe contener los términos CV correspondientes al análisis efectuado dentro del elemento <*AnalysisSummary*> que a su vez están asociados al tipo de aproximación analítica, y por ende al dato cuantitativo, que se incluye en el fichero. Junto a la aproximación es necesario incluir el software empleado en el análisis cuantitativo para poder asociar estos valores a los distintos niveles de cuantificación: *features*, péptidos, proteínas y por último grupos de proteínas.

El elemento <*InputFile*> hace referencia a los datos a partir de los cuales se ha llevado a cabo el análisis, incluyendo los datos de espectrometría de masas (por ejemplo formateados como mzML), los datos relativos a la identificación (por ejemplo formateados como mzIdentML), la base de datos de secuencias de aminoácidos

1 http://psidev.cvs.sourceforge.net/viewvc/psidev/psi/psi-ms/mzML/controlledVocabulary/psi-ms.obo

empleada para la identificación y los parámetros de configuración o ficheros de métodos requeridos para el análisis, como puede ser las transiciones para un análisis dirigido o SRM (por ejemplo el formato traML). El formato captura una descripción del software y la versión del mismo en el elemento <*SoftwareList*>. Los pasos del análisis se agrupan en el elemento <*DataProcessingList*> y son descritos utilizando términos CVs. Cualquier referencia bibliográfica se incluye en el elemento <*BibliographicReference*>.

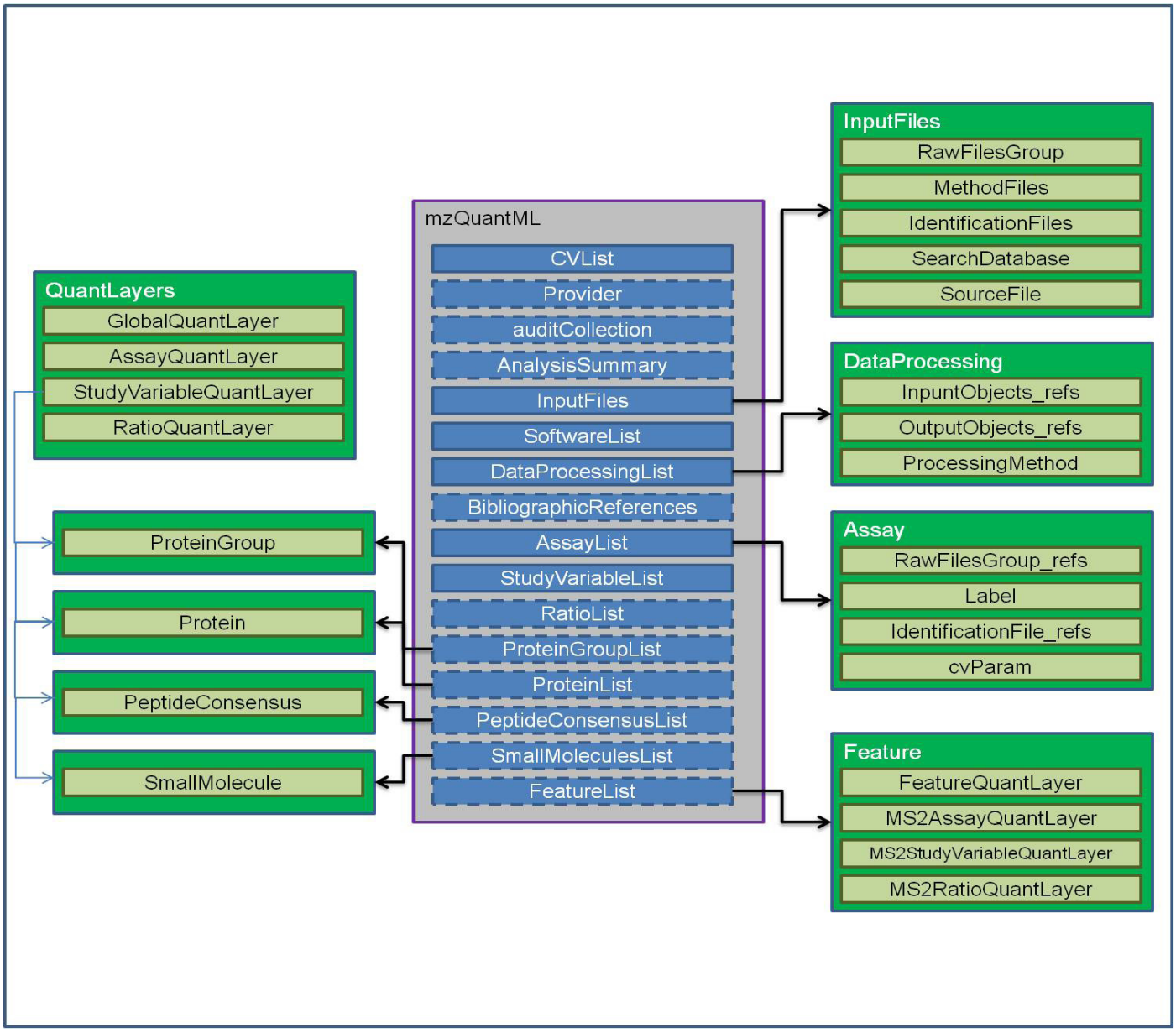

Figura R.1.5: Esquema del modelo mzQuantML. Adaptación de (Walzer, Qi et al. 2013).

Diseño experimental

El diseño experimental se basa en dos elementos clave. El primero <*Assay*> representa el análisis de una muestra biológica sencilla. Los análisis adicionales efectuados sobre la misma muestra o réplicas se modelan como elementos adicionales <*Assay*>. Este elemento es versátil, de tal forma que permite almacenar la información requerida tanto en aquellas técnicas en donde la comparación se lleva a cabo a partir de una única carrera o *run*, como en aquellas en donde el análisis se lleva a cabo a partir de varios *runs*, incluyendo aquellas técnicas basadas en marcaje, donde adicionalmente a los ficheros de espectrometría de masas se debe incluir la etiqueta que diferencia los péptidos ligeros y pesados en aproximaciones de marcaje isotópico como SILAC.

El segundo elemento es <*StudyVariable*> y se utiliza para agrupar un conjunto de elementos <*Assay*> asociados con una misma condición biológica, por ejemplo enferma o control, y por tanto un estado cuantitativo a reportar. Un elemento <*StudyVariable*> contendrá una colección de réplicas biológicas, a partir de las cuales el software de análisis ha calculado un valor (usualmente media aritmética) de los datos cuantitativos.

Durante el desarrollo de mzQuantML, se observó que muchas de las aplicaciones software calculaban un ratio como resultado en lugar de proporcionar las intensidades empleadas en el cálculo de dicho ratio. Éste era empleado indistintamente tanto para representar la relación entre abundancia de proteínas como de péptidos. Esta característica se consideró y se incoporó como caso de uso para la implantación de mzQuantML. Para modelarlo se incluyó el elemento <*RatioList*> que agrupa los distintos elementos <*Ratio*> que pueden aparecer en un análisis cuantitativo.

Cada elemento <*Ratio*> define un numerador y un denominador, cada uno relacionado con un elemento <*StudyVariable*> o un elemento <*Assay*>, dependiendo si el valor a devolver corresponde con la relación entre las condiciones biológicas del estudio o los análisis individuales respectivamente.

Valores cuantitativos en mzQuantML

El formato está basado en una estructura matricial diseñada para que sea flexible y no muy pesada en términos de almacenamiento. Esta estructura está incluida en el elemento <*QuantLayer*> y consiste en una matriz bidimensional de valores numéricos. Existen distintos subtipos del elemento <*QuantLayer*> y se nombran en relación a la capa de experimento desde donde se extraen los valores: Experimento o <*Assay*>, Condición biológica o <*Study variable*>, Ratio o <*Ratio*> y así sucesivamente en cada uno de los elemento con valores cuantitativos a reportar. Estos elementos forman las columnas de la matriz.

La localización del elemento <*QuantLayer*> dentro del fichero define el tipo de objeto al que pertenecen: grupos de proteínas, proteínas, péptidos o *features*, que forman las filas de la matriz. Por ejemplo el elemento <*AssayQuantLayer*> dentro de <*ProteinList*> contendrá una matriz donde las columnas corresponderán con los elementos <*Assay*> o experimentos en donde las proteínas <*Protein*> han sido detectadas y que estarán representadas a lo largo de las filas de la matriz.

En el material suplementario se han incluido diferentes ficheros mzQuantML atendiendo a las diferentes técnicas que se han incluido en el formato estándar.

R2: Establecer nuevos métodos y herramientas que permitan una visión integral de los experimentos basados en espectrometría de masas y anotados según los estándares internacionales

Como se ha establecido en los resultados de la sección R1, un experimento proteómico puede dividirse en varias etapas o fases que al combinarse explican la relación existente entre una realidad biológica y su correspondiente evidencia analítica. Dependiendo del objetivo final del experimento, se puede optar por diferentes enfoques, combinando incluso diferentes técnicas, para lograr el resultado, ya bien sea para separar una muestra compleja, la identificación de péptidos y proteínas o la cuantificación de estos.

Igualmente se han descrito los estándares de datos que existen para tratar de forma uniforme los datos provenientes de cada una de las fases anteriores, e incluso el formato PRIDE XML que engloba en su esquema información relativa a las fases de espectrometría de masas e identificación de péptidos y proteínas. Incluso, con la incorporación del software PRIDESpotMapper (Medina-Aunon, Kenyani et al. 2011) se podría introducir la separación de proteínas siempre y cuando se haya efectuado mediante electroforesis.

Sin embargo, el esquema PRIDE XML presenta principalmente dos inconvenientes: primero, no es un formato cómodo para usuarios finales debido a la complejidad intrínseca de estar basado en XML, como la que se deriva de la relación de los elementos XML en el esquema principal; y segundo, requiere cierto

procesamiento adicional en la traducción de los datos binarios incluidos en el fichero, como por ejemplo, datos de espectrometría de masas para poder chequear los datos contenidos. El trabajo incluido en esta tesis doctoral, que resuelve estos problemas hacía el usuario final, supuso el primer entorno gráfico en donde la información correspondiente a un experimento proteómico se mostraba de forma agrupada a la vez que permitiría de una forma sencilla de validar los resultados (Fig. R2.1).

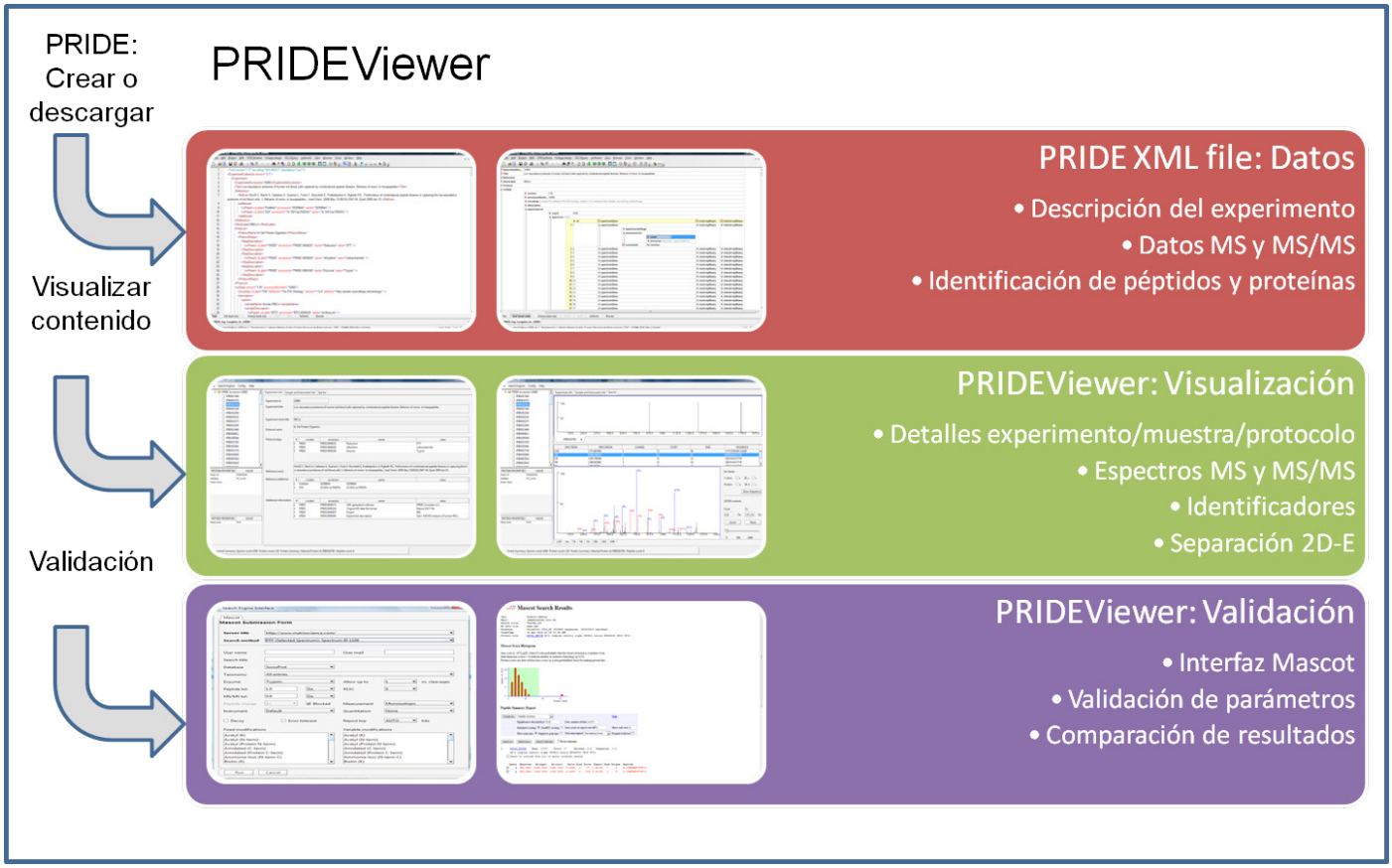

Figura R2.1: Principales características de PRIDEViewer. Adaptación de (Medina-Aunon, Carazo et al. 2011).

PRIDEViewer

PRIDEViewer (Medina-Aunon, Carazo et al. 2011) es una herramienta de ejecución local (*standalone*) desarrollada en Java que permite al usuario trazar completamente el flujo de trabajo proteómico cuyo principal objetivo sea la identificación de péptidos y proteínas. Sin embargo esta herramienta no incluye ninguna ayuda en relación a cálculos cuantitativos, consecuencia que cuando se desarrolló PRIDEViewer, el formato PRIDE no incorporaba internamente este tipo de experimentos.

La presentación de la información sigue la división que ofrece en el esquema PRIDE a través de sus secciones principales:

a) Proteínas identificadas.
b) Descripción del experimento.
c) Descripción de la plataforma analítica empleada, fundamentalmente espectrometría de masas.
d) Espectros de masas MS y tándem MS/MS.
e) Información adicional de péptidos y proteínas identificadas.

La integración de estas secciones permite llevar a cabo una rápida inspección visual de los resultados, haciendo de PRIDEViewer una herramienta muy útil para la revisión de los experimentos proteómicos, tanto por parte de los proveedores o experimentalistas que han realizado el experimento, como por usuarios ajenos, como pueden ser revisores externos, cuyo objetivo es certificar los resultados reportados. Al tratarse de un aspecto esencial en la filosofía de esta aplicación, a continuación se presentan algunas de las ventajas que se aportan en relación a la verificación de resultados.

Visualización y emparejamiento de espectros tándem MS/MS y péptidos

Un usuario experimentado puede analizar con cierta seguridad la calidad de un emparejamiento entre un espectro MS/MS y su correspondiente secuencia peptídica simplemente por la representación gráfica de dicho emparejamiento. Esta característica es controlada por la herramienta PRIDEViewer gracias a la presentación gráfica entre la secuencia del péptido asociada a su identificación y el correspondiente espectro MS/MS, mostrando las series de fragmentación usuales b e y. A esta interpretación del espectro a partir de la secuencia identificada hay que sumar las series de fragmentación complementarias como son a – x y c – z. Los datos mostrados proporcionan una útil y rápida interpretación de los principales elementos relacionados con una proteína o péptido. Si estos datos se han reportado de forma correcta dentro del fichero PRIDE XML, la herramienta PRIDEViewer los leerá, interpretará y mostrará correctamente, permitiendo a cualquier usuario comparar el emparejamiento entre el espectro teórico (que se obtiene a partir de la secuencia de aminoácidos) y los picos (iones fragmento) contenidos en el fichero de resultados (Fig. R.2.2).

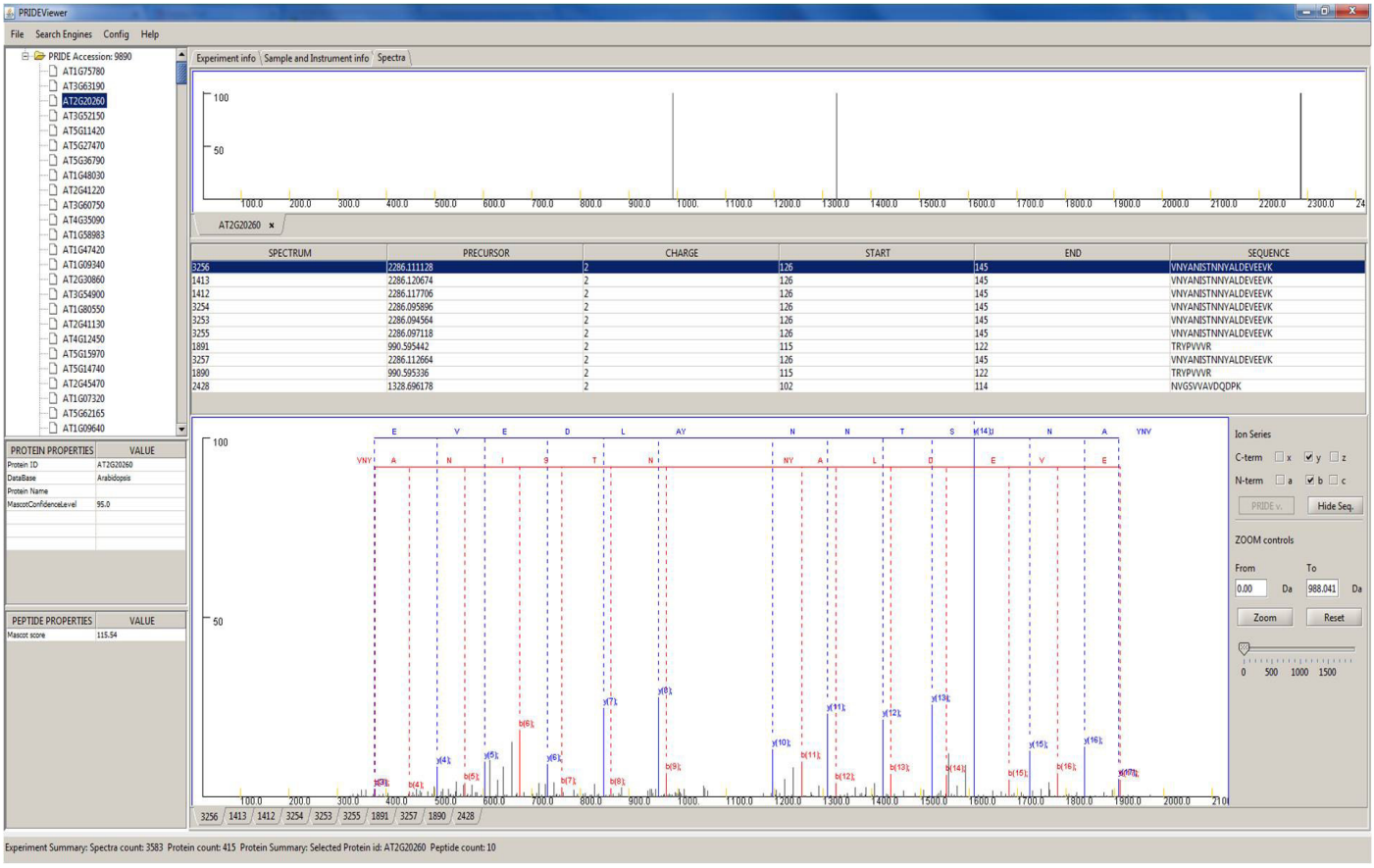

Figura R2.2: Captura de pantalla de la interfaz de usuario de PRIDEViewer.

Complementariamente a la representación gráfica del espectro, se incluye información adicional relativa a cada proteína y/o péptido como: 1) base de datos de secuencias utilizada en la identificación, 2) parámetros relativos a la identificación como la puntuación del candidato (proteína o péptido) o el umbral de confianza,

3) secuencia de la proteína identificada 4) PTMs detectadas en la secuencia del péptido o 5) anotaciones como son las enfermedades o rutas metabólicas relacionadas, función, la localización celular. Este conjunto de elementos permite en una simple representación obtener información valiosa a la hora de validar la ejecución del experimento.

Interface al motor de búsqueda MASCOT

Aunque PRIDEViewer es una valiosa ayuda a la hora de visualizar la información relativa al experimento que se ha reportado, no existe una forma sencilla y automática para validar los datos de espectrometría de masas contenidos en el fichero PRIDE XML. Este hecho es consecuencia de que PRIDE XML no sea un formato habitual como entrada en los motores de búsqueda. Por lo tanto, no es factible repetir el proceso de búsqueda, tal y como ha sido descrito en el fichero XML y comprobar si, se obtienen los mismos resultados en términos de identificación de péptidos y/o proteínas.

En este sentido, se ha desarrolló una interface gráfica (GUI) para el motor de búsqueda Mascot (www.matrixscience.com/mascot) que permite enviar, en tiempo de ejecución, los datos de espectrometría de masas reportados en el fichero PRIDE XML y validarlos de manera inmediata.

Para permitir un rango más amplio de comparaciones, se han implementado todos los modos de envío que Mascot contempla. Esto permite tanto el reanálisis conjunto de todos los espectros relacionados con una determinada proteína (PMF o PMF+MS/MS) como una evaluación atómica para cada espectro MS/MS asociado con un único péptido identificado.

En relación a la forma en que los parámetros de búsqueda son cargados en el formulario, PRIDEViewer descarga en tiempo de ejecución los párametros y opciones disponible (p.e. las bases de datos de secuencias), desde el servidor, asegurando que dicha información esté actualizada. En relación al servidor a conectar, las opciones de configuración permiten las consultas tanto el servidor público (http://www.matrixscience.com) como en servidores locales.

Envío por lotes

Independientemente de la evaluación de las proteínas que se puede llevar a cabo gracias a la GUI de Mascot descrita en el punto anterior, en multitud de ocasiones es necesario una profunda y completa revisión de todo el conjunto de datos contenidos en un fichero PRIDE XML. Por tanto, el envío y posterior comprobación de cada una de las proteínas presentes no es un método sistemático que pueda aplicarse cuando el número de proteínas incluidas en el fichero de resultados es superior a unas cuantas decenas.

Por este motivo se han incluido en la herramienta dos módulos que permitan una revisión global de los datos reportados. El primero –*Mascot Batch Search*- automáticamente evalúa la información relacionada con cada una de las proteínas de forma individualizada. Es este sentido, todos los datos de espectrometría de masas relacionados con un determinada proteína se agrupan y son enviados de manera conjunta como un único envío. Para poder efectuarlo de manera global, se genera una cola, en donde cada elemento a enviar a Mascot será la proteína en cuestión junto con todos los espectros que permiten respaldar su identificación positiva.

Una vez que todos los envíos han finalizado, cada fila de la tabla de los resultados mostrará el primer candidato que al repetir el proceso de identificación se ha obtenido. Cabe esperar que este dato coincida con la proteína que está siendo evaluada en la mayoría de las ocasiones. Si no fuese el caso, la herramienta localiza la proteína evaluada entre todas las identificadas, especificando la posición dentro de esta lista. Con ambas utilidades, se permite un rápido chequeo de los datos reportados.

El segundo módulo permite un envío completo de todos los espectros contenidos en el fichero PRIDE XML sin considerar el vínculo de estos con los péptidos y/o proteínas identificados. Este módulo permite envíos con un número elevado de espectros y permite obtener una vista general, en cuestiones de número de proteínas, de la información reportada.

R3: Establecer nuevos métodos para la extracción de información a partir de bases de datos biológicas y el conjunto de proteínas identificadas en un experimento

Tal y como se establece en la introducción, son dos los enfoques que tradicionalmente se han seguido para poder llevar a cabo la tarea de extracción de la información de una forma metódica. La primera es la minería de texto o *text mining* (Hearst 1999) y tiene como objetivo examinar un conjunto de documentos escritos en lenguaje natural, y por tanto no estructurados según un lenguaje formal y la segunda es la denominada recuperación de información o *information retrieval* (Cowie and Lehnert 1996) y se basa en el almacenamiento, organización y acceso a información que si está estructurada y etiquetada o catalogada. En un contexto proteómico, son múltiples las fuentes que pueden ampliar el conocimiento biológico que se desprende a partir de un conjunto de proteínas dentificadas en un experimento. Esta información permite vincular las condiciones experimentales de la muestra analizada, como puede ser una cierta patología, con las anotaciones funcionales que explican el rol biológico que dichas proteínas desempeñan.

Para poder mejorar el panorama que se presenta en un contexto proteómico, se ha desarrollado la herramienta PIKE: *Protein Information and Knowledge Extractor* (Medina-Aunon, Paradela et al. 2010). Esta herramienta permite enlazar aquellas proteínas que se han identificado gracias al uso de las bases de datos de secuencias

con los principales repositorios públicos de información biológica, funcional y patológica de una manera sencilla y transparente para el usuario. Junto a esta importante faceta, PIKE además permite exportar los resultados utilizando el formato de datos PRIDE XML para su posterior análisis por otras herramientas.

Aplicación PIKE

La herramienta PIKE está desarrollada íntegramente en Java y se apoya en el modelo cliente-servidor (*thin-client*) multicapa para su ejecución. El cliente es un navegador web y se encarga sólo de la generación de la consulta inicial según las preferencias del usuario y una vez terminado el proceso de la presentación de los resultados. Por su parte el servidor lleva a cabo toda la operación de la consulta, realizando las funciones típicas de las capas de negocio y almacenamiento.

Independientemente de otros parámetros como puede ser el tipo de identificador utilizado o la información a recuperar, el principal dato de entrada es un conjunto de proteínas identificadas o de interés en el contexto de un experimento proteómico. Estas deben reunirse en un fichero de entrada que permite dos formatos. El primer formato es texto plano y dentro de éste cada fila representa una proteína caracterizada por su identificador dentro de la base de datos de secuencias utilizada en la fase de identificación. El segundo es el formato es PRIDE XML.

Identificadores de las bases de datos

Uno de los hitos principales a cubrir en la aplicación PIKE es la posibilidad de utilizar diversos identificadores de bases de datos de secuencias como entradas permitidas en las listas de proteínas que inician el proceso. Esto permite ampliar el número de escenarios posibles de cara a los usuarios finales. Sin embargo, esta característica requiere una comprobación adicional para asegurar la perfecta correspondencia entre los identificadores utilizados y las fuentes de datos.

Este paso, que consiste en un elemento central sobre el que descarga el algoritmo de recuperación de PIKE, no incorpora bases de datos internas de referencias cruzadas que requieran una actualización constante en el lado del servidor, sino que hace uso de los recursos disponibles que existen por Internet como el ofrecido en el sitio web del *Protein Information Resource* (PIR) (http://pir.georgetown.edu/).

Algoritmo PIKE

El algoritmo principal de PIKE está dividido en tres módulos (Fig. R3.1). Estos módulos permiten realizar una división de las tareas acorde a la estructura multicapa de la aplicación.

El primer módulo o administrador del flujo de trabajo (*WFMM: Workflow Manager Module*) comprueba la validez de los parámetros y el fichero de entrada que contiene las proteínas y selecciona las fuentes de información más idóneas relacionadas con la consulta efectuada por el usuario.

El segundo módulo, recuperación de información (*IRM: Information Retrieval Module*), establece los pasos necesarios para capturar la información de las fuentes seleccionadas en el módulo anterior. Dentro de este

módulo se comprueba que estas fuentes seleccionadas están correctamente conectadas para permitir una recuperación secuencial de la información para cada proteína de la lista inicial. En cada uno de los pasos establecidos, el módulo selecciona el método que mejor se ajuste a la disponibilidad y accesibilidad del dato de interés a recuperar.

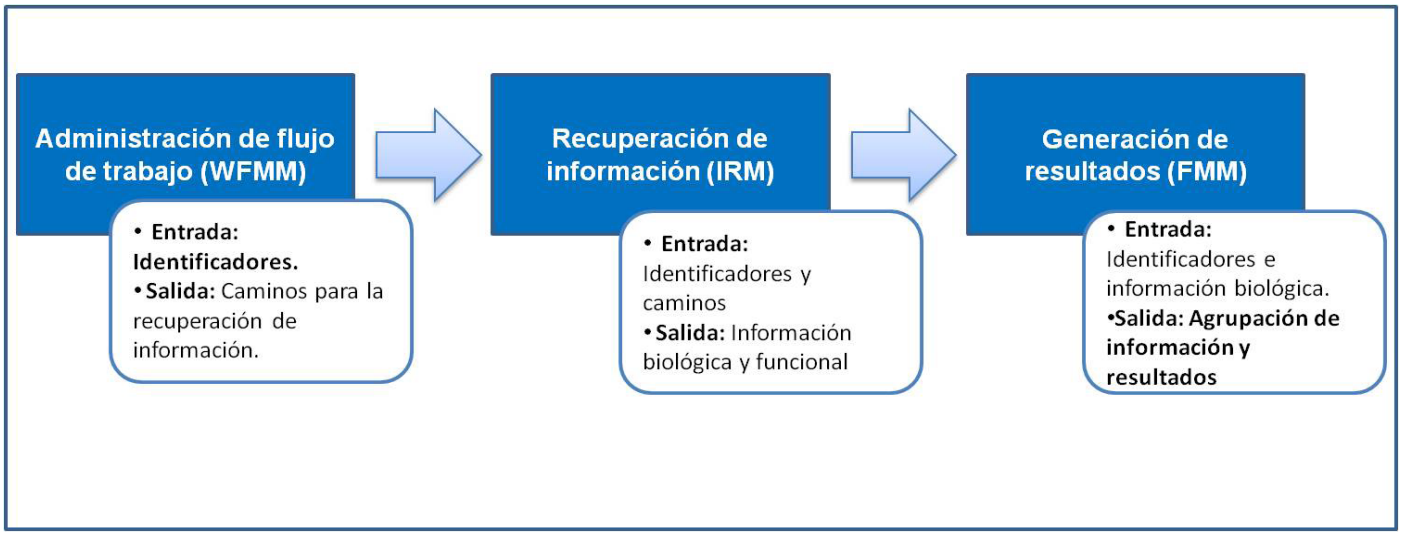

Figura R3.1: Estructura de funcionamiento de la aplicación PIKE. Los distintos módulos van recopilando la información necesaria para proporcionar las anotaciones de interés para el usuario.

Por último el tercer módulo (*FMM: File Manager Module)* reúne toda la información que se asocia con el listado inicial de proteínas y acorde a los parámetros de búsqueda introducidos en la consulta con el fin de generar los ficheros de resultados. En este sentido, PIKE genera resultados en diferentes formatos, incluyendo HTML, ficheros separados por comas (CSV: *Comma separated values*), ficheros de texto o distribuciones gráficas en JPEG (*Joint Photographic Experts Group*) o SVG (*Scalable Vector Graphics*).

Extracción de conocimiento

A pesar de que la recopilación de información significa una valiosa fuente de conocimiento, si esta información no está ordenada y/o categorizada puede resultar abrumadora e inútil. Es por ello, que aunque la salida por defecto de la herramienta PIKE consiste en una única tabla donde se resume toda la información solicitada por el usuario, se incluyen otra serie de documentos y ficheros donde dicha información es agrupada por categorías (Fig. R3.2). Este es el caso concreto de los ficheros separados por comas (CSV) que se generan en paralelo y agrupan las proteínas de interés atendiendo a su clasificación por la ontología Gene Ontology para cada una de sus tres categorías, las rutas metabólicas ofrecidas según el repositorio KEGG (*Kyoto Encyclopedia of Genes and Genomes*) o la relación con enfermedades y procesos patológicos según la librería OMIM (*Online Mendelian Inheritance in Man*). De manera complementaria los agrupamientos que se llevan a cabo a partir de los términos Gene Ontology son representados gráficamente para obtener una visión global la relaciones jerárquicas que se establecen a través de la ontología.

Enriquecimiento de la información relativa a proteínas

El enriquecimiento permite determinar la relevancia de la sub o sobre expresión de un determinado conjunto de proteínas en comparación con un conjunto de referencia (p.e. proteoma humano). Las anotaciones recuperadas pueden poner de manifiesto que el resultado de un experimento contiene un porcentaje significativo de todas las proteínas del conjunto de referencia que compartan una determinada función o estén

asociadas a un proceso biológico o patología concreta. Con el fin de incorporar una valoración estadística en términos de enriquecimiento, PIKE incorpora un enlace dinámico con la herramienta DAVID (*Database for Annotation, Visualization and Integrated Discovery*) (Dennis, Sherman et al. 2003), que proporciona una medida estadística suficientemente contrastada sobre las relaciones que se establecen entre las anotaciones de GO, KEGG y OMIM.

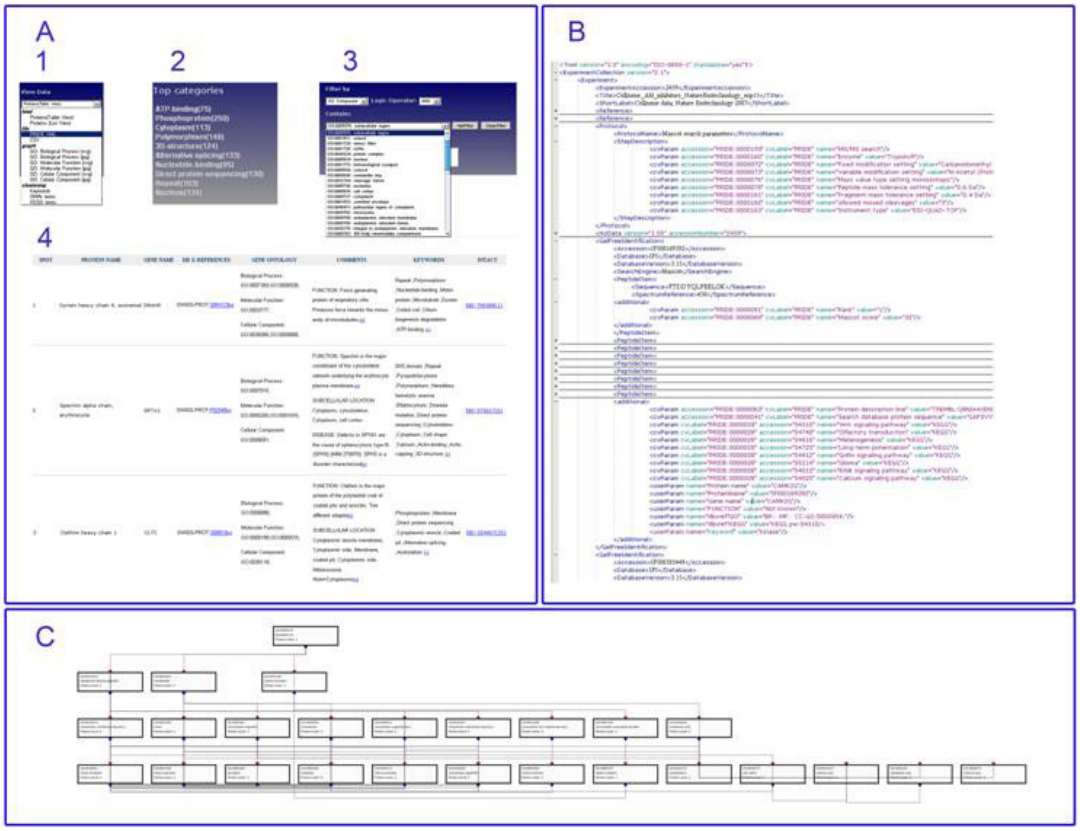

Figura R3.2: Capturas de pantalla de las algunas vistas de resultados que ofrece PIKE. A. Interfaz HTML resumiendo toda la información recuperada y con capacidad para hacer filtros sencillos. B. Resultados en formato PRIDE XML para su integración en otras herramientas. C. Vista gráfica del agrupamiento según Gene Ontology. Fuente: (Medina-Aunon, Paradela et al. 2010) .

PIKE en ejecución

Son cuatro los ejemplos que demuestran la mejora que ofrece PIKE a la hora de analizar un conjunto de datos teniendo en cuenta el contexto biológico al que las anotaciones dirigen. Estos cuatro ejemplos han sido seleccionados para representar diferentes escenarios de recuperación de información completa y no redundante a partir de las bases de datos públicas.

Análisis de un subproteoma de complejidad media

En este primer caso de uso se deseaba comprobar la eficacia de PIKE en relación a las anotaciones recuperadas manualmente de 57 proteínas humanas reportadas en el trabajo de Paradela et. al (Paradela, Bravo et al. 2005).Toda la información recopilada manualmente por los autores fue recuperada automáticamente por PIKE, mostrando una eficacia similar en la detección de los puntos de interés que los autores resaltaban en la publicación. Es importante destacar que PIKE mejoró los resultados incluidos en este trabajo debido fundamentalmente a dos aspectos. El primero era el evidente ahorro de tiempo, reduciendo el tiempo de

búsqueda manual de varios días a unos pocos minutos. El segundo era el acceso a los datos actualizados, lo que derivó en encontrar características adicionales que no fueron reportadas en la publicación y que mejoraba por tanto el análisis de las anotaciones.

Análisis del glicoproteoma de suero humano

El siguiente ejemplo demuestra como la herramienta PIKE puede integrarse en un flujo de análisis, concretamente con su incorporación en la plataforma software ProteinScape™. En total 22 carreras LC-MS/MS fueron reunidas y enfrentadas a los motores de búsqueda Mascot y Phenyx para la identificación de 352 proteínas únicas. Esta lista fue revisada manualmente y tras la eliminación de proteínas contaminantes, la lista se redujo a 250 glicoproteínas. Toda los datos experimentales se encontraban almacenados de forma permanente en el sistema de administración de experimentos ProteinScape™. A través de una interfaz gráfica (GUI) desarrollada sobre esta plataforma se pudo enviar estas 250 proteínas a PIKE y obtener las anotaciones relativas a la ontología Gene Ontology (GO) y las palabras clave (*Keywords*) de la base de datos UniProt. Los resultados fueron tratados por Microsoft® Excel y revelaron que un 22% de las proteínas enviadas estaban involucradas en respuesta inmune, un 20% relacionadas con el metabolismo, un 10% en transporte de moléculas y por último un 10% en comunicación celular. En relación a la localización celular de las proteínas, un 53% se ubicaban fuera de la célula, 26% en la membrana celular y por último un 20% en el citosol. Estos resultados demostraron que, tras la integración en ProteinScape™, PIKE mantiene su eficiencia, resultando una combinación muy útil para análisis de experimentos proteómicos complejos.

Proteínas de plasma humano desde PRIDE

Este tercer ejemplo permite reevaluar y enriquecer los resultados previamente depositados en el repositorio PRIDE. Para este propósito, el experimento más complejo que existía en dicho repositorio cuando se elaboró el artículo fue descargado y analizado por PIKE. Concretamente se optó por un experimento enviado por la iniciativa de Plasma de la organización HUPO (*HUPO Protein Plasma Project:* HPPP) cuyo identificador en el repositorio público de PRIDE era el número 65. Este proyecto comprendía 6071 proteínas y tras su envío a PIKE, las anotaciones recuperadas fueron incorporadas al fichero PRIDE XML, complementado el fichero inicial con la información biológica y funcional disponible. Todo el proceso se llevó a cabo en un tiempo inferior a 6 horas (una media de 1000 proteínas por hora). Una vez la ejecución de PIKE concluyó, las proteínas fueron clasificadas según los criterios introducidos en la búsqueda.

Esa clasificación no arrojó datos muy significativos o relevantes a primera vista, ya que la distribución de anotaciones era muy dispersa. La anotación con mayor frecuencia, polimorfismo, superaba por poco el 5% de las proteínas enviadas. Sin embargo, un análisis más en profundidad determinó que existía un porcentaje muy alto de proteínas redundantes (en torno al 89%). En este caso, la proteína complemento C3 aparecía en 271 ocasiones con diferentes nombres, la Inmonuglobulina A1 225 y la albúmina de suero 146 respectivamente. La eliminación de estas redundancias en el fichero original PRIDE XML es una tarea muy laboriosa y prácticamente inabordable. Sin embargo, los agrupamientos que ofrece PIKE como salidas complementarias en formato CSV permitió eliminar estas rápidamente y repetir el proceso, pero ya con proteínas únicas. Esta reducción supuso bajar de 6071 proteínas a 665 diferenciables según la descripción asociada a cada entrada y un análisis de anotaciones más dirigido y exacto. Como último paso en el proceso, se generó un nuevo PRIDE XML pero conteniendo sólo las proteínas únicas.

Este último ejemplo, contenido en la colaboración liderada por Mario A. Rodríguez-Pérez (Rodriguez-Perez, Medina-Aunon et al. 2008), supuso el uso de PIKE en un contexto más clínico que los ejemplos anteriores y en combinación con un repositorio de gran relevancia como es el atlas de proteínas humanas (HPA: *Human Protein Atlas*) (Uhlen, Bjorling et al. 2005; Uhlen, Oksvold et al. 2010).

En concreto, HPA fue utilizado para obtener distintos listados de proteínas relacionados con distintas etapas clínicas de la enfermedad objeto de estudio. Estos listados a su vez fueron combinados con los biomarcadores descritos en la literatura para esta patología y enviados a PIKE. La información inicial fue complementada con las anotaciones disponibles relacionadas con el role funcional que cada proteína desarrolla, las referencias patológicas proporcionadas por la librería OMIM y la especificidad del tejido.

Este análisis bioinformático determinó de manera automática que 10 proteínas presentaban características de potenciales biomarcadores. Este subconjunto fue contrastado y validado mediante búsqueda bibliográfica. Además, las proteínas sobre expresadas y prognósticas de la patología estudiada presentaban características y funciones habituales en este tipo estudios como son unión, polimorfismos o PTMs concretas como fosforilizaciones. Las restantes 141 proteínas candidatas no pudieron ser validadas como potenciales biomarcadores.

Discusión

Para facilitar su compresión, la discusión se estructura en dos secciones. En primer lugar, se facilita un análisis de los trabajos incluidos en esta memoria de tesis doctoral en una organización similar a los objetivos propuestos. Aquí se discute resumidamente la contribución individual de cada uno de ellos. A continuación, se procede a una discusión general de la aportación que esta tesis doctoral ofrece a la comunidad proteómica en particular desde una óptica computacional.

D1.1. Estudio e implantación de estándares internacionales para la representación de datos según las diferentes fases que componen los experimentos proteómicos

Los trabajos reunidos ofrecen a los usuarios proteómicos nuevos formatos estándares y herramientas bioinformáticas. Estos representan una mejora sustancial en la estandarización e integración de los datos derivados de experimentos en el flujo de trabajo cotidiano. Como ya se comentó en la introducción, los estándares de datos se diseñaron con el objetivo de facilitar la colaboración y el intercambio de estos entre laboratorios con independencia de la plataforma analítica empleada (fundamentalmente datos generados por los espectrómetros de masas). Sin embargo, también suponen una mejora en términos de eficiencia, al permitir a los investigadores centrarse en el análisis de los datos en lugar de dedicar recursos a tareas repetitivas como la lectura y detección de errores en los ficheros empleados habitualmente como entradas y/o salidas de los paquetes software, comprensión de los formatos de intercambio o la localización y extracción de la información relevante.

Entre los trabajos recopilados en esta tesis doctoral se ofrecen dos nuevos estándares XML para las fases de separación por electroforesis y cuantificación de péptidos y proteínas. Son de destacar dos características

muy relevantes relacionadas con la definición de ambos formatos. Primeramente, estos han sido fruto de una abierta colaboración entre miembros de HUPO-PSI de diferentes nacionalidades, incluyendo a expertos en los campos proteómicos y computacionales. Este hecho proporciona a estos formatos un alto nivel de consenso, confiriéndoles la catalogación de estándares HUPO-PSI. La segunda es que son formatos vivos en el sentido de que están abiertos a actualizaciones y mejoras que incluyan nuevos casos de uso que surjan en el futuro. Para dicho fin, se han dispuesto foros de discusión, como listas de correo y repositorios abiertos (*Google Code*), y la frecuente celebración de reuniones entre sus desarrolladores y miembros de la comunidad científica decididos a involucrarse.

El primero de estos formatos, el estándar GelML (Gibson, Hoogland et al. 2010) captura la información necesaria relacionada con experimentos basados en electroforesis, tanto para los protocolos tradicionales 1-DE, 2-DE y DIGE, como usos no tan habituales. La información asociada contiene un alto componente de anotación manual, lo que la hace más susceptible a la aparición de errores. GelML permite que dicha información sea comprobada y validada a la hora de registrar un experimento.

El estándar mzQuantML (Walzer, Qi et al. 2013) por su parte presenta una estructura versátil para incluir los datos derivados de un experimento de proteómica cuantitativa. Como se ha descrito en los resultados, esta área es extremadamente compleja y puede generar ingentes volúmenes de datos que incluyen entre otros la definición del experimento, las condiciones iniciales, el protocolo de cuantificación o los resultados. Debido a esta complejidad, mzQuantML se ha diseñado tanto para cubrir las diferentes aproximaciones metodológicas existentes como nuevos escenarios y técnicas que puedan aparecer en el futuro. Además incluye elementos que permite mejorar las herramientas software, incluyendo opciones de visualización y de re-análisis de los datos. Por último asegura que los datos a enviar a los repositorios públicos contienen la información necesaria para poder realizar un seguimiento completo del cálculo de los resultados finales en términos de cuantificación de péptidos y/o proteínas.

Sin embargo, un elemento negativo que siempre ha acompañado la definición de estándares en proteómica es la incorporación de estos en el trabajo rutinario. Generalizando, los esquemas XML suelen ser bastante extensos y en muchas ocasiones la información no se presenta de forma compacta o dividida en bloques que facilitaría su tratamiento. Se requiere entonces editores o visualizadores creados *ex profeso*, pero que por lo habitual no son desarrollados hasta que el formato alcanza cierta madurez. Este hecho origina un retraso, a veces considerable, entre la publicación del estándar y su puesta en funcionamiento en entornos reales, demorando la detección de errores del formato en este contexto.

Más en detalle, son dos las principales razones principales que limitan la extensión del estándar GelML. La primera es la falta de herramientas que incorporen métodos para exportar o importar ficheros GelML, consecuencia de que la mayor parte del software de análisis de imagen es software propietario. A diferencia del consenso alcanzado con los estándares mzML (Martens, Chambers et al. 2011) y mzIdentML (Jones, Eisenacher et al. 2012) y los principales fabricantes y desarrolladores de software especializado, un aspecto que claramente ha perjudicado en la incorporación de GelML en software propietario fue la ausencia de colaboración por parte de las compañías líderes en este sector en las sesiones y eventos celebrados para la definición del estándar.

El segundo factor determinante es su lenta adaptación al flujo de trabajo. La electroforesis es un proceso laborioso formado por un conjunto de protocolos ejecutados de forma manual sin proporcionar resultados tratables por métodos computacionales como sí lo son las relativas a identificación y cuantificación de proteínas.

A pesar de estos dos características limitantes, GelML se incluye como formato de salida del entorno de trabajo ProteoRed Miape Web ToolKit (PMWKT) (Medina-Aunon, Martinez-Bartolome et al. 2011), lo que permite a los usuarios extraer todos los pasos realizados en la separación por electroforesis a partir de los experimentos almacenados en la ProteoRed Miape Database (PMDB) (Martinez-Bartolome, Medina-Aunon et al. 2010), para su posterior análisis

En relación a mzQuantML son varios los proyectos de código abierto (Maxquant (Cox and Mann 2008)) Proteosuite (Gonzalez-Galarza, Lawless et al. 2012) OpenMS y TOPP (Bertsch, Gropl et al. 2011)) que incluyen este formato como elemento basal en el análisis cuantitativo. Sin embargo, su versatilidad es un factor limitante para aquellos usos en los que únicamente interesan los resultados finales. Para estos casos se ha desarrollado otro estándar (mzTab (http://www.psidev.info/mztab)) que simplemente resume el conjunto mínimo de datos que permite evaluar los resultados de un experimento.

En resumen, ambos formatos suponen una notable contribución al mundo académico en términos de anotación y reporte de experimentos, pero su uso está limitado en función del número de herramientas que incorporen estos estándares y el nivel de obligatoriedad a la hora de publicar los resultados. Por tanto, si en el futuro son integrados en las herramientas habituales de tratamiento y análisis de datos, con total transparencia de cara al usuario, su uso será incuestionable.

Por otro lado, dos nuevos entornos de trabajo se han implementado para fomentar la integración entre los estándares HUPO-PSI, las guías MIAPE y el formato PRIDE XML. Además, ambos han sido diseñados para su incorporación en el flujo rutinario del laboratorio tanto por su integración con otras herramientas como en su orientación para usuarios finales y no sólo para aquellos más especializados en bioinformática.

En este sentido, el primero de ellos, PRIDESpotMapper (Medina-Aunon, Kenyani et al. 2011) supone la única herramienta disponible que permite anotar como PRIDE XML los miles de experimentos basados en separación por electroforesis que se publican cada año y, más importante, su posterior envío al repositorio PRIDE para su difusión entre la comunidad científica. Hasta la fecha de publicación de este trabajo, ningún experimento basado en electroforesis había sido correctamente depositado en este repositorio de referencia, impidiendo que una parte significativa de los experimentos publicados en proteómica pudiesen ser compartidos o reanalizados. La anotación según el formato PRIDE XML de experimentos basados en electroforesis, cumple con todas las especificaciones sintácticas y semánticas del esquema PRIDE, resultando apto para su posterior envío a la base de datos que almacena este tipo de ficheros.

El segundo, PMWTK, supone un entorno computacional más ambicioso en el sentido que permite a los usuarios la conexión y sincronización de formatos XML de HUPO PSI, guías MIAPE y formato PRIDE XML de una forma sencilla y eficaz. Este entorno permite ahorrar a los investigadores una cantidad significativa de tiempo y esfuerzo a la hora de cumplir con las reglas de publicación de las principales revistas proteómicas. Además proporciona un sistema de validación asegurando que tanto los ficheros XML como las guías MIAPE del mismo experimento se complementen, validen y conecten automáticamente, proporcionando una sintaxis válida y correcta de las anotaciones de los datos experimentales.

Aunque ambos marcos de trabajo presentan características únicas en el tratamiento e integración de los diferentes estándares de datos, facilitando así un salto cualitativo en la compartición de datos experimentales, requieren un mayor acercamiento al usuario final. Ambas herramientas deben abstraer los datos realmente relevantes de toda la información que se incluye en los ficheros de resultados para poder examinarlos

y reanalizarlos de una forma más liviana. Además, un segundo factor a considerar es evitar al usuario la obligación de conocer la localización de todos los ficheros necesarios para generar los informes y/o resultados finales. En este sentido, la generación de un fichero válido según el esquema PRIDE XML y que contenga los datos relativos a una separación por electroforesis requiere un arduo trabajo manual de mapeo de la localización de los spots dentro del gel y sus respectivas identificaciones. Este mismo problema es extrapolable al PMWTK cuando es empleado dentro del flujo de trabajo rutinario en un laboratorio. En este contexto, PMWTK necesita trabajar localmente con multitud de ficheros que además presentan el inconveniente de ser difícilmente manejables por su tamaño.

Un aspecto común a ambas herramientas es que tienen como objetivo final la generación de resultados formateados según PRIDE XML. En este sentido, los marcos de trabajo descritos no son los únicos disponibles, puesto que otra herramienta, PRIDE Converter, permite realizar la misma tarea. La última versión de esta herramienta, PRIDE Converter 2 (Cote, Griss et al. 2012) presenta ciertas ventajas sobre los entornos descritos. Los principales son la integración de más formatos de salida de espectrómetros de masas y motores de búsqueda como ficheros de entada o la inclusión del formato mzTab para proporcionar un primer vínculo entre PRIDE XML y los experimentos de proteómica cuantitativa. Sin embargo, no permite incorporar toda la información exigida por las guías de publicación MIAPE de una forma metódica como si lo hace PMWTK. Este hecho hace de PMWTK una herramienta idónea para el cumplimiento de las directrices de publicación de las revistas especializadas en proteómica.

D1.2. Establecer nuevos métodos y herramientas que permitan una visión integral de los experimentos basados en espectrometría de masas y anotados según los estándares internacionales

Como se ha descrito a lo largo de esta memoria, el repositorio PRIDE supone un pilar básico entre la comunidad científica para el intercambio de resultados. Sin embargo, al basarse en un formato de intercambio XML, no resulta trivial la interpretación de los datos contenidos por usuarios no expertos y sin la ayuda de paquetes de software especializados en la edición de este tipo de ficheros.

La herramienta PRIDEViewer (Medina-Aunon, Carazo et al. 2011) supuso la primera iniciativa que acerca la integración de datos que propone el formato PRIDE XML al usuario final, representado en una interfaz amigable toda la información contenida en un fichero de resultados. Datos relativos a la aproximación metodológica, separación 2D-E, espectrometría de masas, identificación de péptidos y proteínas y por último anotaciones funcionales asociadas a dichas identificaciones son mostradas al usuario de forma totalmente accesible y navegable. Este entorno permite en un primer término una visualización completa de los datos contenidos y aporta al usuario final ciertas utilidades que facilitan la comprobación de los resultados rápida y visualmente. Ejemplo de esta característica es la funcionalidad que permite emparejar el péptido identificado con el espectro de fragmentación correspondiente. Esta vista permite, de forma casi instantánea, una primera aproximación para evaluar la calidad del emparejamiento. Además permite comprobar la asignación de fragmentos no sólo para las series principales de fragmentación (b e y), sino también con las secundarias (a, c, x y z).

Esta primera vista gráfica si bien resulta muy útil e intuitiva para usuarios proteómicos, no es suficiente cuando el número de péptidos y proteínas a validar es considerable. En este sentido, un elemento que es

único hasta el momento ofrece PRIDEViewer de forma exclusiva es la validación masiva de resultados. En estos casos, la interfaz con el motor de búsqueda Mascot permite de una manera global comprobar si las proteínas descritas en los resultados han sido correctamente identificadas. A partir de los péptidos asociados a cada proteína, y por extensión con los espectros MS/MS vinculados con dichos péptidos, los datos primeramente se envían al motor de búsqueda y a continuación se comparan con los nuevos resultados, proporcionando una vía rápida y sencilla de validación. Estas utilidades hacen de PRIDEViewer una herramienta muy accesible para los usuarios no familiarizados con el formato PRIDE XML para obtener una vista global y validada de los resultados.

Posteriormente a PRIDEViewer se desarrolló la herramienta PRIDE Inspector (Wang, Fabregat et al. 2012) que si bien mejoraba algunos aspectos de PRIDEViewer, no se incluyeron otros que se pueden entender como elementales. Algunas de las ventajas que PRIDE Inspector ofrece en comparación con PRIDEViewer son: la capacidad para trabajar con ficheros PRIDE XML sin limitación de tamaño, conexión directa con el repositorio PRIDE y la generación de una serie de representaciones gráficas que proporcionan una rápida inspección visual de los resultados como son las distribuciones de número de péptidos por proteína, número de picos por espectro MS/MS, o intensidades de los picos, entre otros. Sin embargo un elemento ausente en PRIDE Inspector es la validación de resultados. En resumen y aunque la explotación de la resultados podría ser mayor, PRIDEViewer si ofrece un entorno adecuado para la validación de experimentos, permitiendo a revisores y personal dedicado a la verificación de datos una solución muy orientada a sus necesidades.

D1.3. Establecer nuevos métodos para la extracción de información a partir de bases de datos biológicas y el conjunto de proteínas identificadas en un experimento

La extracción de información en proteómica supone un salto cualitativo en el entendimiento del rol funcional que desempeña un determinado conjunto de proteínas. Gracias, por un lado, a la información recopilada en las bases de datos en las últimas décadas y, por otro, a las herramientas que permiten un tratamiento automático de la información, esta tarea es actualmente abordable.

PIKE (Medina-Aunon, Paradela et al. 2010) es una de estas herramientas y cuenta con la peculiaridad de que ha sido específicamente diseñada para la recuperación de información de experimentos proteómicos. Los resultados incluidos en esta memoria demuestran que PIKE vincula sin ambigüedad la información disponible en las bases de datos con un conjunto de proteínas previamente identificadas. A partir de los identificadores de estas proteínas, PIKE encuentra el camino correcto para recuperar la información biológica y funcional disponible en las bases de datos que resulta de interés para el usuario. Además PIKE es la única herramienta que hasta el momento ofrece una vista global de los resultados, resumiendo la información más relevante en múltiples vistas, incluyendo representaciones gráficas que agrupan las proteínas clasificadas según las anotaciones recuperadas.

Una de las principales ventajas de PIKE es su capacidad de resolver los dos problemas básicos, y desgraciadamente habituales, derivados del uso de los códigos de acceso de las proteínas de distintas bases de datos. En contraposición con otras herramientas (Zeeberg, Feng et al. 2003; Al-Shahrour, Minguez et al. 2005; Carmona-Saez, Chagoyen et al. 2007), PIKE ofrece un amplio listado de códigos para utilizar como entradas (UniprotKB -*Accession* e *Identification*-, EBI IPI, NCBI nr –*gi y refeq*-, GenBank, EMBL, PDB,

DDBJ y Entrez Gene ID) que abarca los más empleados a la hora de proporcionar un listado de proteínas identificadas. En segundo lugar, establece las referencias cruzadas entre los distintos códigos de acceso, lo que permite acceder a la información requerida independientemente del código de entrada utilizado.

Otra característica a destacar de esta herramienta es su adaptación dinámica. En lugar de almacenar las anotaciones a proporcionar en una base de datos local, ofrece un sistema configurable basado en diferentes módulos que permite acceder en tiempo real a la última versión disponible de las anotaciones. A pesar que el proceso se ralentiza y es dependiente de la conexión a Internet, ofrece la doble ventaja de asegurar que la información mostrada está actualizada a la vez que ahorra esfuerzos derivados de mantener una base de datos interna. Una consecuencia de este dinamismo es la capacidad de actualización y adaptación que presenta la herramienta ante nuevas situaciones sin tener que alterar significativamente la lógica de la aplicación. Un ejemplo de esta filosofía es la integración con DAVID (*Database for Annotation, Visualization and Integrated Discovery*). De esta forma la aplicación no sobrecarga internamente la operativa con complejos cálculos estadísticos y traslada estos a una herramienta suficientemente contrastada y experta en estas tareas. Resumidamente, la aplicación hace uso de recursos disponibles, descentralizando el trabajo computacional a realizar.

Por último y fruto del compromiso con las corrientes de estandarización, PIKE incluye el formato PRIDE XML como entrada y/o salida estándar. Esto permite compartir y enriquecer resultados previamente recogidos en este formato facilitando su incorporación en el flujo habitual de análisis de datos. Para que esta integración sea efectiva, PIKE inserta las anotaciones recopiladas utilizando los vocabularios controlados (CVs) disponibles para que su posterior procesamiento se ejecute sin ambigüedad y permita un análisis preciso de la información. Categorías como Gene Ontology, IntAct, KEGG y OMIM disponen de términos reservados en los CVs para este fin. Una vez que un resultado es exportado como PRIDE XML y tras la anotación y validación según los CVs, el resultado enriquecido con las anotaciones puede enviarse al repositorio de referencia PRIDE.

En resumen, y tal como se ha descrito en la sección de resultados, PIKE es una herramienta perfectamente aplicable a estudios proteómicos independientemente de su tamaño (desde unas pocas decenas de proteínas hasta miles) y el enfoque (enriquecimiento, eliminación de redundancias o agrupamiento de proteínas según anotaciones) de estos.

D2.1. Discusión final

En la sección anterior se describe resumidamente como los trabajos incluidos en esta memoria de tesis doctoral aportan individualmente diversas mejoras en el campo de la proteómica. Tanto en la definición y diseño de nuevos formatos de intercambio como en el desarrollo de herramientas, se contribuye a resolver problemas específicos de esta área. Sin embargo la complementariedad de las soluciones recopiladas permite ofrecer una óptica integradora y global.

Todos los trabajos se enmarcan en la mejora completa del flujo de trabajo en proteómica. Desde la preparación de la muestra hasta el procesado, análisis, validación y enriquecimiento de los datos generados, los resultados descritos resuelven carencias notables. En primer lugar, en el registro de información dentro del campo de la separación de la muestra y más en concreto las limitaciones existentes en la difusión de resultados, finales e intermedios, cuando se emplea la separación por electroforesis. En relación a la fase de análisis por espectrometría de masas y posterior identificación de péptidos y proteínas, se aportan soluciones que permiten tanto su vínculo con los datos obtenidos en la fase de separación, en el caso de empleo de electroforesis, como la visualización y validación de los resultados de espectrometría. Por último, y en pro de lograr una mejor compresión del contexto biológico de un experimento, es evidente que el resultado de un experimento hay que trasladarlo al plano funcional. En este sentido, y partiendo del resultado final del experimento proteómico, (p.e. un conjunto de identificadores de proteínas), se propone una herramienta expresamente diseñada e implementada para dicho fin.

Todos los datos generados en el desarrollo de un experimento que reúna todas o parte de estas fases son perfectamente conectados en base a los trabajos recopilados en esta tesis doctoral. En concreto, gracias a la incorporación del formato PRIDE XML en todos las herramientas descritas, es posible incorporar todos detalles relevantes que permita la trazabilidad completa del experimento.

Siguiendo con el flujo de trabajo en proteómica, otra mejora es la capacidad de normalizar y estandarizar los resultados cuantitativos. Este aspecto, si bien no se integra obligatoriamente en el flujo clásico de separación, análisis e identificación de péptidos y proteínas, es una tendencia cada vez más frecuente por su potencial tanto en la investigación básica como en su aplicación clínica. La comparación entre distintas condiciones fisiopatológicas (p.e control o sana versus problema o enfermo) y su traslación al diseño experimental en un contexto proteómico presenta una extraordinaria relevancia en el campo de la investigación actual. Con la inclusión de un estándar versátil que sea capaz de tratar con los datos generados, se facilita una vía que permite el análisis en detalle de estos datos.

Otro factor fundamental es la difusión e intercambio de información. Tanto el empleo de estándares de datos *per se*, como la integración de estos en herramientas bioinformáticas permite por defecto que cualquier resultado pueda ser reinterpretado y complementado desde una plataforma de análisis distinta a la original. Además las tendencias actuales en el campo fomentan la colaboración y trabajo en red en lugar de centralizar los recursos en un único lugar con suficiente potencial para poder liderar proyectos de investigación de gran envergadura. La asociación de laboratorios que prestan servicios de investigación básica es actualmente una realidad. Un ejemplo es la Red Carlos III-ProteoRed (Paradela, Escuredo et al. 2006) que establece un marco colaborativo entre más de 20 laboratorios geográficamente distribuidos por la Península Ibérica. Fruto de esta colaboración es la sinergia que se establecen entre los participantes, donde cada laboratorio aporta

su experiencia en determinadas áreas y/o tecnologías que sirva de base para la investigación española en proteómica. Desde el punto de vista computacional, uno de los mayores compromisos adquiridos es el uso de los estándares definidos por HUPO PSI para el intercambio de información entre los laboratorios participantes (Martinez-Bartolome, Blanco et al. 2010). Este compromiso ha resultado clave para el desarrollo herramientas bioinformáticas (incluidas en esta memoria entre otras) resaltando la necesidad de analizar los datos generados en un contexto más amplio que el laboratorio proveedor.

Otro ejemplo, éste en el plano internacional, es la caracterización del proteoma humano (HPP – *Human Proteome Project*) (http://www.thehpp.org/) (Legrain, Aebersold et al. 2011). El objetivo de este proyecto es generar un mapa con las proteínas codificadas en los aproximadamente 20.300 genes del genoma humano para posteriormente usarlo como recurso en el entendimiento de su rol funcional y mejorar el diagnóstico y tratamiento de enfermedades. Para lograr este objetivo, el problema global se ha dividido en equipos de trabajo, donde cada uno tiene asignado el conjunto de proteínas codificado en un cromosoma humano concreto.

Entre las obligaciones que los participantes en esta iniciativa mundial asumen en cuestión de tratamiento de datos, está el envío de los datos MS/MS generados al repositorio de referencia PRIDE. Este hecho exige emplear, en primer lugar, los estándares de datos para representar los resultados obtenidos y en segundo lugar, emplear las herramientas basadas en estos estándares para analizar dichos resultados.

Las razones de esta obligación están basadas en la configuración en red del HPP. Cada equipo involucrado en la generación del mapa de proteínas humanas debe compartir sus resultados con el resto de participantes para permitir reanalizar los datos generados. La única forma de asegurar que este proceso se lleve a cabo correctamente es que los resultados sean formateados según un estándar adecuado.

En concreto, la participación española en este proyecto (consorcio spHPP), se centra en las proteínas asociadas al cromosoma 16. El consorcio se apoya en la infraestructura de ProteoRed para poder realizar los experimentos que permitan cumplir con objetivos fijados. En este sentido y como parte activa del proyecto, algunas de las herramientas descritas en esta memoria de tesis doctoral han sido empleadas para establecer un marco de trabajo bioinformático común que cumpla con los requerimientos de difundir y compartir los resultados analíticos obtenidos (Segura, Medina-Aunon et al. 2013). En detalle, y considerando el flujo de trabajo establecido, es necesario en primer término recopilar todos los resultados de los laboratorios que forman el consorcio spHPP, formatearlos según los estándares de datos y emplear PMWTK para vincular con las guías de publicación MIAPE. Una vez que los datos son almacenados en la base de datos (PMDB) y tras un análisis conjunto de todas las aportaciones, son formateados según PRIDE XML para su posterior envío al repositorio PRIDE. Aunque tanto la generación del fichero de resultados PRIDE XML como su posterior envío al repositorio se podría llevar a cabo a partir de la conexión de otras herramientas existentes, el uso de PMWTK incorpora la información requerida en las guías de publicación MIAPE a los resultados.

Adicionalmente, y entre la generación del PRIDE XML y el envío al repositorio PRIDE, es posible enriquecer los resultados con el uso de PIKE, para completar el nivel de anotación de los péptidos y proteínas resultantes. Un ejemplo de esto último lo encontramos en el reciente trabajo de otro equipo del proyecto HPP, el cromosoma 19 liderado por Suecia (Nilsson, Berven et al. 2013). En este caso PIKE se empleó primeramente para obtener un agrupamiento de las proteínas identificadas tanto por su función biológica y localización celular. En segundo lugar para localizar las proteínas anotadas como fosforiladas.

En resumen, los trabajos recopilados en esta memoria de tesis doctoral complementan el flujo completo de trabajo proteómico desde la óptica del tratamiento bioinformático integral de los datos y resultados. La inserción de los estándares de datos, tanto los ya existentes como los incluidos en esta memoria, en las herramientas desarrolladas permite una integración más ambiciosa de las distintas variaciones que se pueden dar en el flujo de análisis de datos. Además ofrecen un marco adecuado a las tendencias actuales de trabajo en red y mantienen el compromiso con la estandarización.

Conclusiones

1. Como parte del grupo de trabajo HUPO-PSI, se ha diseñado un nuevo formato estándar (GelML) basado en XML para recopilar y vincular todos los datos generados en una separación por electroforesis. El formato soporta tanto el uso habitual de electroforesis 1D-E, 2D-E y DIGE como variaciones particulares no tan frecuentes.

2. Como parte del grupo de trabajo HUPO-PSI, se ha diseñado un nuevo formato estándar (mzQuantML) basado en XML para recopilar y vincular todos los datos derivados de un experimento de proteómica cuantitativa. El formato admite la mayoría de las aproximaciones experimentales para realizar dicha cuantificación y es adaptable a nuevos protocolos que se definan.

3. Se ha diseñado e implementado un nuevo marco de trabajo (ProteoRed MIAPE Web ToolKit PMWTK) para poder vincular y sincronizar los estándares de datos propuestos por HUPO-PSI, las guías MIAPE y el formato PRIDE XML. El marco de trabajo PMWTK ha servido de base para poder analizar globalmente las aportaciones del consorcio español (spHPP) al proyecto HPP.

4. Se ha diseñado e implementado un marco de trabajo (PRIDESpotMapper) para almacenar resultados derivados de experimentos basados en separación por electroforesis empleando el formato PRIDE XML. PRIDESpotMapper permitió el primer envío completo al repositorio PRIDE de un experimento basado en separación por electroforesis.

5. Se ha diseñado e implementado una herramienta que permite visualizar gráfica y amigablemente toda la información contenida en un fichero PRIDE XML. Esta herramienta permite realizar una validación de los datos relativos a espectrometría de masas e identificación de péptidos y proteínas incluidos en el fichero PRIDE XML.

6. Se ha diseñado e implementado un servicio de búsqueda de anotaciones funcionales para un conjunto de proteínas resultantes de un experimento. El sistema está específicamente orientado a experimentos proteómicos resolviendo problemas propios de las técnicas empleadas en dichos experimentos.

Referencias

Abecasis, G. R., D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles and G. A. McVean (2010). "A map of human genome variation from population-scale sequencing." Nature **467**(7319): 1061-1073.

Aebersold, R. and D. R. Goodlett (2001). "Mass spectrometry in proteomics." Chem Rev **101**(2): 269-295.

Al-Shahrour, F., R. Diaz-Uriarte and J. Dopazo (2004). "FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes." Bioinformatics **20**(4): 578-580.

Al-Shahrour, F., P. Minguez, J. M. Vaquerizas, L. Conde and J. Dopazo (2005). "BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments." Nucleic Acids Res **33**(Web Server issue): W460-464.

Ambihapathy, K., T. Yalcin, H.-W. Leung and A. G. Harrison (1997). "Pathways to Immonium Ions in the Fragmentation of Protonated Peptides." Journal of Mass Spectrometry **32**(2): 209-215.

Anderson, L. and C. L. Hunter (2006). "Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins." Mol Cell Proteomics **5**(4): 573-588.

Anderson, N. G. and N. L. Anderson (1996). "Twenty years of two-dimensional electrophoresis: past, present and future." Electrophoresis **17**(3): 443-453.

Anderson, N. L. and N. G. Anderson (2002). "The human plasma proteome: history, character, and diagnostic prospects." Mol Cell Proteomics **1**(11): 845-867.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Barsnes, H., J. A. Vizcaino, I. Eidhammer and L. Martens (2009). "PRIDE Converter: making proteomics data-sharing

easy." Nat Biotechnol **27**(7): 598-599.

Benson, D. A., M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and E. W. Sayers (2013). "GenBank." Nucleic Acids Res **41**(Database issue): D36-42.

Bertsch, A., C. Gropl, K. Reinert and O. Kohlbacher (2011). "OpenMS and TOPP: open source software for LC-MS data analysis." Methods Mol Biol **696**: 353-367.

Biemann, K. (1990). "Appendix 5. Nomenclature for peptide fragment ions (positive ions)." Methods Enzymol **193**: 886-887.

Biemann, K. (1992). "Mass spectrometry of peptides and proteins." Annu Rev Biochem **61**: 977-1010.

Biemann, K. and S. A. Martin (1987). "Mass spectrometric determination of the amino acid sequence of peptides and proteins." Mass Spectrometry Reviews **6**(1): 1-75.

Binz, P. A., R. Barkovich, R. C. Beavis, D. Creasy, D. M. Horn, R. K. Julian, Jr., S. L. Seymour, C. F. Taylor and Y. Vandenbrouck (2008). "Guidelines for reporting the use of mass spectrometry informatics in proteomics." Nat Biotechnol **26**(8): 862.

Birney, E., J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, M. S. Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. A. Greenbaum, R. M. Andrews, P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhami, S. C. Dillon, M. O. Dorschner, H. Fiegler, P. G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K. D. James, B. E. Johnson, E. M. Johnson, T. T. Frum, E. R. Rosenzweig, N. Karnani, K. Lee, G. C. Lefebvre, P. A. Navas, F. Neri, S. C. Parker, P. J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox, M. Yu, F. S. Collins, J. Dekker, J. D. Lieb, T. D. Tullius, G. E. Crawford, S. Sunyaev, W. S. Noble, I. Dunham, F. Denoeud, A. Reymond, P. Kapranov, J. Rozowsky, D. Zheng, R. Castelo, A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I. L. Hofacker, R. Baertsch, D. Keefe, S. Dike, J. Cheng, H. A. Hirsch, E. A. Sekinger, J. Lagarde, J. F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermuller, J. Hertel, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korbel, O. Emanuelsson, J. S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M. C. Dickson, D. J. Thomas, M. T. Weirauch, J. Gilbert, J. Drenkow, I. Bell, X. Zhao, K. G. Srinivasan, W. K. Sung, H. S. Ooi, K. P. Chiu, S. Foissac, T. Alioto, M. Brent, L. Pachter, M. L. Tress, A. Valencia, S. W. Choo, C. Y. Choo, C. Ucla, C. Manzano, C. Wyss, E. Cheung, T. G. Clark, J. B. Brown, M. Ganesh, S. Patel, H. Tammana, J. Chrast, C. N. Henrichsen, C. Kai, J. Kawai, U. Nagalakshmi, J. Wu, Z. Lian, J. Lian, P. Newburger, X. Zhang, P. Bickel, J. S. Mattick, P. Carninci, Y. Hayashizaki, S. Weissman, T. Hubbard, R. M. Myers, J. Rogers, P. F. Stadler, T. M. Lowe, C. L. Wei, Y. Ruan, K. Struhl, M. Gerstein, S. E. Antonarakis, Y. Fu, E. D. Green, U. Karaoz, A. Siepel, J. Taylor, L. A. Liefer, K. A. Wetterstrand, P. J. Good, E. A. Feingold, M. S. Guyer, G. M. Cooper, G. Asimenos, C. N. Dewey, M. Hou, S. Nikolaev, J. I. Montoya-Burgos, A. Loytynoja, S. Whelan, F. Pardi, T. Massingham, H. Huang, N. R. Zhang, I. Holmes, J. C. Mullikin, A. Ureta-Vidal, B. Paten, M. Seringhaus, D. Church, K. Rosenbloom, W. J. Kent, E. A. Stone, S. Batzoglou, N. Goldman, R. C. Hardison, D. Haussler, W. Miller, A. Sidow, N. D. Trinklein, Z. D. Zhang, L. Barrera, R. Stuart, D. C. King, A. Ameur, S. Enroth, M. C. Bieda, J. Kim, A. A. Bhinge, N. Jiang, J. Liu, F. Yao, V. B. Vega, C. W. Lee, P. Ng, A. Shahab, A. Yang, Z. Moqtaderi, Z. Zhu, X. Xu, S. Squazzo, M. J. Oberley, D. Inman, M. A. Singer, T. A. Richmond, K. J. Munn, A. Rada-Iglesias, O. Wallerman, J. Komorowski, J. C. Fowler, P. Couttet, A. W. Bruce, O. M. Dovey, P. D. Ellis, C. F. Langford, D. A. Nix, G. Euskirchen, S. Hartman, A. E. Urban, P. Kraus, S. Van Calcar, N. Heintzman, T. H. Kim, K. Wang, C. Qu, G. Hon, R. Luna, C. K. Glass, M. G. Rosenfeld, S. F. Aldred, S. J. Cooper, A. Halees, J. M. Lin, H. P. Shulha, X. Zhang, M. Xu, J. N. Haidar, Y. Yu, Y. Ruan, V. R. Iyer, R. D. Green, C. Wadelius, P. J. Farnham, B. Ren, R. A. Harte, A. S. Hinrichs, H. Trumbower, H. Clawson, J. Hillman-Jackson, A. S. Zweig, K. Smith, A. Thakkapallayil, G. Barber, R. M. Kuhn, D. Karolchik, L. Armengol, C. P. Bird, P. I. de Bakker, A. D. Kern, N. Lopez-Bigas, J. D. Martin, B. E. Stranger, A. Woodroffe, E. Davydov, A. Dimas, E. Eyras, I. B. Hallgrimsdottir, J. Huppert, M. C. Zody, G. R. Abecasis, X. Estivill, G. G. Bouffard, X. Guan, N. F. Hansen, J. R. Idol, V. V. Maduro, B. Maskeri, J. C. McDowell, M. Park, P. J. Thomas, A. C. Young, R. W. Blakesley, D. M. Muzny, E. Sodergren, D. A. Wheeler, K. C. Worley, H. Jiang, G. M. Weinstock, R. A. Gibbs, T. Graves, R. Fulton, E. R. Mardis, R. K. Wilson, M. Clamp, J. Cuff, S. Gnerre, D. B. Jaffe, J. L. Chang, K. Lindblad-Toh, E. S. Lander, M. Koriabine, M. Nefedov, K. Osoegawa, Y. Yoshinaga, B. Zhu and P. J. de Jong (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146): 799-816.

Buckingham, S. (2004). "Bioinformatics: data's future shock." Nature **428**(6984): 774-777.

Carlos Setubal, J. M. (1997). Introduction to Computational Molecular Biology. Boston, PWS publishing company.

Carmona-Saez, P., M. Chagoyen, F. Tirado, J. M. Carazo and A. Pascual-Montano (2007). "GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists." Genome Biol **8**(1): R3.

Carr, S., R. Aebersold, M. Baldwin, A. Burlingame, K. Clauser and A. Nesvizhskii (2004). "The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data." Mol Cell Proteomics **3**(6): 531-533.

Chamrad, D. C., G. Korting, K. Stuhler, H. E. Meyer, J. Klose and M. Bluggel (2004). "Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data." Proteomics **4**(3): 619-628.

Cifuentes, A. (2013). "Foodomics. Advance Mass Spectrometry in Modern Food Science and Nutrition" Willey.

Clauser, K. R., P. Baker and A. L. Burlingame (1999). "Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching." Anal Chem **71**(14): 2871-2882.

Claverie, J. M. (2001). "Gene number. What if there are only 30,000 human genes?" Science **291**(5507): 1255-1257.

Cleland, W. W. (1964). "Dithiothreitol, a New Protective Reagent for Sh Groups." Biochemistry **3**: 480-482.

Collins, F. S. (2001). "Contemplating the end of the beginning." Genome Res **11**(5): 641-643.

Consortium, T. C. S. a. A. (2005). "Initial sequence of the chimpanzee genome and comparison with the human genome." Nature **437**(7055): 69-87.

Cote, R. G., J. Griss, J. A. Dianes, R. Wang, J. C. Wright, H. W. van den Toorn, B. van Breukelen, A. J. Heck, N. Hulstaert, L. Martens, F. Reisinger, A. Csordas, D. Ovelleiro, Y. Perez-Rivevol, H. Barsnes, H. Hermjakob and J. A. Vizcaino (2012). "The PRoteomics IDEntification (PRIDE) Converter 2 framework: an improved suite of tools to facilitate data submission to the PRIDE database and the ProteomeXchange consortium." Mol Cell Proteomics **11**(12): 1682-1689.

Cowie, J. and W. Lehnert (1996). "Information extraction." Commun. ACM **39**(1): 80--91.

Cox, J. and M. Mann (2008). "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification." Nat Biotechnol **26**(12): 1367-1372.

Csordas, A., D. Ovelleiro, R. Wang, J. M. Foster, D. Rios, J. A. Vizcaino and H. Hermjakob (2012). "PRIDE: quality control in a proteomics data repository." Database (Oxford) **2012**: bas004.

Delwiche, F. A. (2008). "Searching MEDLINE via PubMed." Clin Lab Sci **21**(1): 35-41.

Dennis, G., Jr., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane and R. A. Lempicki (2003). "DAVID: Database for Annotation, Visualization, and Integrated Discovery." Genome Biol **4**(5): P3.

Deutsch, E. W., M. Chambers, S. Neumann, F. Levander, P. A. Binz, J. Shofstahl, D. S. Campbell, L. Mendoza, D. Ovelleiro, K. Helsens, L. Martens, R. Aebersold, R. L. Moritz and M. Y. Brusniak (2012). "TraML--a standard format for exchange of selected reaction monitoring transition lists." Mol Cell Proteomics **11**(4): R111 015040.

Domann, P. J., S. Akashi, C. Barbas, L. Huang, W. Lau, C. Legido-Quigley, S. McClean, C. Neususs, D. Perrett, M. Quaglia, E. Rapp, L. Smallshaw, N. W. Smith, W. F. Smyth and C. F. Taylor (2010). "Guidelines for reporting the use of capillary electrophoresis in proteomics." Nat Biotechnol **28**(7): 654-655.

Dong, M. W. (1992). "Tryptic mapping by reversed phase liquid chromatography." Adv Chromatogr **32**: 21-51.

Fazekas de St Groth, S., R. G. Webster and A. Datyner (1963). “Two new staining procedures for quantitative estimation of proteins on electrophoretic strips.” Biochim Biophys Acta **71**: 377-391.

Fenn, J. B., M. Mann, C. K. Meng, S. F. Wong and C. M. Whitehouse (1989). “Electrospray ionization for mass spectrometry of large biomolecules.” Science **246**(4926): 64-71.

Flicek, P., I. Ahmed, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, C. Garcia-Giron, L. Gordon, T. Hourlier, S. Hunt, T. Juettemann, A. K. Kahari, S. Keenan, M. Komorowska, E. Kulesha, I. Longden, T. Maurel, W. M. McLaren, M. Muffato, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, G. R. Ritchie, M. Ruffier, M. Schuster, D. Sheppard, D. Sobral, K. Taylor, A. Thormann, S. Trevanion, S. White, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, J. Harrow, J. Herrero, T. J. Hubbard, N. Johnson, R. Kinsella, A. Parker, G. Spudich, A. Yates, A. Zadissa and S. M. Searle (2013). “Ensembl 2013.” Nucleic Acids Res **41**(Database issue): D48-55.

Garwood, K., T. McLaughlin, C. Garwood, S. Joens, N. Morrison, C. F. Taylor, K. Carroll, C. Evans, A. D. Whetton, S. Hart, D. Stead, Z. Yin, A. J. Brown, A. Hesketh, K. Chater, L. Hansson, M. Mewissen, P. Ghazal, J. Howard, K. S. Lilley, S. J. Gaskell, A. Brass, S. J. Hubbard, S. G. Oliver and N. W. Paton (2004). “PEDRo: a database for storing, searching and disseminating experimental proteomics data.” BMC Genomics **5**: 68.

Gibson, F., L. Anderson, G. Babnigg, M. Baker, M. Berth, P. A. Binz, A. Borthwick, P. Cash, B. W. Day, D. B. Friedman, D. Garland, H. B. Gutstein, C. Hoogland, N. A. Jones, A. Khan, J. Klose, A. I. Lamond, P. F. Lemkin, K. S. Lilley, J. Minden, N. J. Morris, N. W. Paton, M. R. Pisano, J. E. Prime, T. Rabilloud, D. A. Stead, C. F. Taylor, H. Voshol, A. Wipat and A. R. Jones (2008). “Guidelines for reporting the use of gel electrophoresis in proteomics.” Nat Biotechnol **26**(8): 863-864.

Gibson, F., C. Hoogland, S. Martinez-Bartolome, J. A. Medina-Aunon, J. P. Albar, G. Babnigg, A. Wipat, H. Hermjakob, J. S. Almeida, R. Stanislaus, N. W. Paton and A. R. Jones (2010). “The gel electrophoresis markup language (GelML) from the Proteomics Standards Initiative.” Proteomics **10**(17): 3073-3081.

Gonzalez-Galarza, F. F., C. Lawless, S. J. Hubbard, J. Fan, C. Bessant, H. Hermjakob and A. R. Jones (2012). “A Critical Appraisal of Techniques, Software Packages, and Standards for Quantitative Proteomic Analysis.” OMICS **(9)**: 431-42. .

Gygi, S. P., B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb and R. Aebersold (1999). “Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.” Nat Biotechnol **17**(10): 994-999.

Hearst, M. A. (1999). Untangling text data mining. Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. College Park, Maryland, Association for Computational Linguistics**:** 3-10.

Henson, J., G. Tischler and Z. Ning (2012). “Next-generation sequencing and large genome assemblies.” Pharmacogenomics **13**(8): 901-915.

Holcapek, M., R. Jirasko and M. Lisa (2012). “Recent developments in liquid chromatography-mass spectrometry and related techniques.” J Chromatogr A **1259**: 3-15.

Hoogland, C., M. O’Gorman, P. Bogard, F. Gibson, M. Berth, S. J. Cockell, A. Ekefjard, O. Forsstrom-Olsson, A. Kapferer, M. Nilsson, S. Martinez-Bartolome, J. P. Albar, S. Echevarria-Zomeno, M. Martinez-Gomariz, J. Joets, P. A. Binz, C. F. Taylor, A. Dowsey and A. R. Jones (2010). “Guidelines for reporting the use of gel image informatics in proteomics.” Nat Biotechnol **28**(7): 655-656.

Horth, P., C. A. Miller, T. Preckel and C. Wenz (2006). “Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis.” Mol Cell Proteomics **5**(10): 1968-1974.

Hortin, G. L. and D. Sviridov (2010). “The dynamic range problem in the analysis of the plasma proteome.” J Proteomics **73**(3): 629-636.

Huang, T., J. Wang, W. Yu and Z. He (2012). "Protein inference: a review." Brief Bioinform **13**(5): 586-614.

Jones, A. R., K. Carroll, D. Knight, K. Maclellan, P. J. Domann, C. Legido-Quigley, L. Huang, L. Smallshaw, H. Mirzaei, J. Shofstahl and N. W. Paton (2010). "Guidelines for reporting the use of column chromatography in proteomics." Nat Biotechnol **28**(7): 654.

Jones, A. R., M. Eisenacher, G. Mayer, O. Kohlbacher, J. Siepen, S. Hubbard, J. Selley, B. Searle, J. Shofstahl, S. Seymour, R. Julian, P. A. Binz, E. W. Deutsch, H. Hermjakob, F. Reisinger, J. Griss, J. A. Vizcaino, M. Chambers, A. Pizarro and D. Creasy (2012). "The mzIdentML data standard for mass spectrometry-based proteomics results." Mol Cell Proteomics.

Jones, A. R., M. Miller, R. Aebersold, R. Apweiler, C. A. Ball, A. Brazma, J. Degreef, N. Hardy, H. Hermjakob, S. J. Hubbard, P. Hussey, M. Igra, H. Jenkins, R. K. Julian, Jr., K. Laursen, S. G. Oliver, N. W. Paton, S. A. Sansone, U. Sarkans, C. J. Stoeckert, Jr., C. F. Taylor, P. L. Whetzel, J. A. White, P. Spellman and A. Pizarro (2007). "The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics." Nat Biotechnol **25**(10): 1127-1133.

Jones, P., R. G. Cote, L. Martens, A. F. Quinn, C. F. Taylor, W. Derache, H. Hermjakob and R. Apweiler (2006). "PRIDE: a public repository of protein and peptide identifications for the proteomics community." Nucleic Acids Res **34**(Database issue): D659-663.

Kaiser, J. (2002). "Proteomics. Public-private group maps out initiatives." Science **296**(5569): 827.

Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." Nucleic Acids Res **28**(1): 27-30.

Karas, M. and F. Hillenkamp (1988). "Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons." Anal Chem **60**(20): 2299-2301.

Keshishian, H., T. Addona, M. Burgess, E. Kuhn and S. A. Carr (2007). "Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution." Mol Cell Proteomics **6**(12): 2212-2229.

Khatri, P., S. Draghici, G. C. Ostermeier and S. A. Krawetz (2002). "Profiling gene expression using onto-express." Genomics **79**(2): 266-270.

Kuzyk, M. A., D. Smith, J. Yang, T. J. Cross, A. M. Jackson, D. B. Hardie, N. L. Anderson and C. H. Borchers (2009). "Multiple reaction monitoring-based, multiplexed, absolute quantitation of 45 proteins in human plasma." Mol Cell Proteomics **8**(8): 1860-1877.

Laemmli, U. K. (1970). "Cleavage of structural proteins during the assembly of the head of bacteriophage T4." Nature **227**(5259): 680-685.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer,

G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi and Y. J. Chen (2001). “Initial sequencing and analysis of the human genome.” Nature **409**(6822): 860-921.

Legrain, P., R. Aebersold, A. Archakov, A. Bairoch, K. Bala, L. Beretta, J. Bergeron, C. H. Borchers, G. L. Corthals, C. E. Costello, E. W. Deutsch, B. Domon, W. Hancock, F. He, D. Hochstrasser, G. Marko-Varga, G. H. Salekdeh, S. Sechi, M. Snyder, S. Srivastava, M. Uhlen, C. H. Wu, T. Yamamoto, Y. K. Paik and G. S. Omenn (2011). “The human proteome project: current state and future direction.” Mol Cell Proteomics **10**(7): M111 009993.

Lieber, D. C. (2001). Introduction to Proteomics: Tools for the New Biology, Humana Press.

Malmstrom, J., M. Beck, A. Schmidt, V. Lange, E. W. Deutsch and R. Aebersold (2009). “Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans.” Nature **460**(7256): 762-765.

Mann, M., P. Hojrup and P. Roepstorff (1993). “Use of mass spectrometric molecular weight information to identify proteins in sequence databases.” Biol Mass Spectrom **22**(6): 338-345.

Martens, L. (2011). “Proteomics databases and repositories.” Methods Mol Biol **694**: 213-227.

Martens, L., M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W. H. Tang, A. Rompp, S. Neumann, A. D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P. A. Binz and E. W. Deutsch (2011). “mzML--a community standard for mass spectrometry data.” Mol Cell Proteomics **10**(1): R110 000133.

Martens, L., H. Hermjakob, P. Jones, M. Adamski, C. Taylor, D. States, K. Gevaert, J. Vandekerckhove and R. Apweiler (2005). “PRIDE: the proteomics identifications database.” Proteomics **5**(13): 3537-3545.

Martin, D., C. Brun, E. Remy, P. Mouren, D. Thieffry and B. Jacq (2004). “GOToolBox: functional analysis of gene datasets based on Gene Ontology.” Genome Biol **5**(12): R101.

Martinez-Bartolome, S., F. Blanco and J. P. Albar (2010). “Relevance of proteomics standards for the ProteoRed Spanish organization.” J Proteomics **73**(6): 1061-1066.

Martinez-Bartolome, S., E. W. Deutsch, P. A. Binz, A. R. Jones, M. Eisenacher, G. Mayer, A. Campos, F. Canals, J. J. Bech-Serra, M. Carrascal, M. Gay, A. Paradela, R. Navajas, M. Marcilla, M. L. Hernaez, M. D. Gutierrez-Blazquez, L. F. Velarde, K. Aloria, J. Beaskoetxea, J. A. Medina-Aunon and J. P. Albar (2013). “Guidelines for reporting quantitative mass spectrometry based experiments in proteomics.” J Proteomics.

Martinez-Bartolome, S., J. A. Medina-Aunon, A. R. Jones and J. P. Albar (2010). “Semi-automatic tool to describe, store and compare proteomics experiments based on MIAPE compliant reports.” Proteomics **10**(6): 1256-1260.

McPherson, J. D., M. Marra, L. Hillier, R. H. Waterston, A. Chinwalla, J. Wallis, M. Sekhon, K. Wylie, E. R. Mardis, R. K. Wilson, R. Fulton, T. A. Kucaba, C. Wagner-McPherson, W. B. Barbazuk, S. G. Gregory, S. J. Humphray, L. French, R. S. Evans, G. Bethel, A. Whittaker, J. L. Holden, O. T. McCann, A. Dunham, C. Soderlund, C. E. Scott, D. R. Bentley, G. Schuler, H. C. Chen, W. Jang, E. D. Green, J. R. Idol, V. V. Maduro, K. T. Montgomery, E. Lee, A. Miller, S. Emerling, Kucherlapati, R. Gibbs, S. Scherer, J. H. Gorrell, E. Sodergren, K. Clerc-Blankenburg, P. Tabor, S. Naylor, D. Garcia, P. J. de Jong, J. J. Catanese, N. Nowak, K. Osoegawa, S. Qin, L. Rowen, A.

Madan, M. Dors, L. Hood, B. Trask, C. Friedman, H. Massa, V. G. Cheung, I. R. Kirsch, T. Reid, R. Yonescu, J. Weissenbach, T. Bruls, R. Heilig, E. Branscomb, A. Olsen, N. Doggett, J. F. Cheng, T. Hawkins, R. M. Myers, J. Shang, L. Ramirez, J. Schmutz, O. Velasquez, K. Dixon, N. E. Stone, D. R. Cox, D. Haussler, W. J. Kent, T. Furey, S. Rogic, S. Kennedy, S. Jones, A. Rosenthal, G. Wen, M. Schilhabel, G. Gloeckner, G. Nyakatura, R. Siebert, B. Schlegelberger, J. Korenberg, X. N. Chen, A. Fujiyama, M. Hattori, A. Toyoda, T. Yada, H. S. Park, Y. Sakaki, N. Shimizu, S. Asakawa, K. Kawasaki, T. Sasaki, A. Shintani, A. Shimizu, K. Shibuya, J. Kudoh, S. Minoshima, J. Ramser, P. Seranski, C. Hoff, A. Poustka, R. Reinhardt and H. Lehrach (2001). "A physical map of the human genome." Nature **409**(6822): 934-941.

Medina-Aunon, J. A., J. M. Carazo and J. P. Albar (2011). "PRIDEViewer: a novel user-friendly interface to visualize PRIDE XML files." Proteomics **11**(2): 334-337.

Medina-Aunon, J. A., J. Kenyani, S. Martinez-Bartolome, J. P. Albar, J. M. Wastling and A. R. Jones (2011). "A DIGE study on the effects of salbutamol on the rat muscle proteome - an exemplar of best practice for data sharing in proteomics." BMC Res Notes **4**: 86.

Medina-Aunon, J. A., R. Krishna, F. Ghali, J. P. Albar and A. J. Jones (2013). "A guide for integration of proteomic data standards into laboratory workflows." Proteomics **13**(3-4): 480-492.

Medina-Aunon, J. A., S. Martinez-Bartolome, M. A. Lopez-Garcia, E. Salazar, R. Navajas, A. R. Jones, A. Paradela and J. P. Albar (2011). "The ProteoRed MIAPE web toolkit: a user-friendly framework to connect and share proteomics standards." Mol Cell Proteomics **10**(10): M111 008334.

Medina-Aunon, J. A., A. Paradela, M. Macht, H. Thiele, G. Corthals and J. P. Albar (2010). "Protein Information and Knowledge Extractor: Discovering biological information from proteomics data." Proteomics **10**(18): 3262-3271.

Miguel Valcárcel Cases, A. G. H. (1988). Técnicas Analíticas de Separación. S. A. Editorial Reverté**:** 333.

Montecchi-Palazzi, L., S. Kerrien, F. Reisinger, B. Aranda, A. R. Jones, L. Martens and H. Hermjakob (2009). "The PSI semantic validator: a framework to check MIAPE compliance of proteomics data." Proteomics **9**(22): 5112-5119.

Monteoliva, L., R. Martinez-Lopez, A. Pitarch, M. L. Hernaez, A. Serna, C. Nombela, J. P. Albar and C. Gil (2011). "Quantitative proteome and acidic subproteome profiling of Candida albicans yeast-to-hypha transition." J Proteome Res **10**(2): 502-517.

Murray, S. and D. Watson (1986). "Bis-trifluoromethylbenzoyl derivatives for steroid analysis by gas chromatography electron capture negative ion chemical ionisation mass spectrometry." J Steroid Biochem **25**(2): 255-260.

Nasukawa, T. and T. Nagano (2001). "Text analysis and knowledge mining system." IBM Syst. J. **40**(4): 967-984.

NCBI_Resource_Coordinators (2013). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **41**(Database issue): D8-D20.

Nilsson, C. L., F. Berven, F. Selheim, H. Liu, J. R. Moskal, R. A. Kroes, E. P. Sulman, C. A. Conrad, F. F. Lang, P. E. Andren, A. Nilsson, E. Carlsohn, H. Lilja, J. Malm, D. Fenyo, D. Subramaniyam, X. Wang, M. Gonzales-Gonzales, N. Dasilva, P. Diez, M. Fuentes, A. Vegvari, K. Sjodin, C. Welinder, T. Laurell, T. E. Fehniger, H. Lindberg, M. Rezeli, G. Edula, S. Hober and G. Marko-Varga (2013). "Chromosome 19 annotations with disease speciation: a first report from the Global Research Consortium." J Proteome Res **12**(1): 135-150.

O'Farrell, P. H. (1975). "High resolution two-dimensional electrophoresis of proteins." J Biol Chem **250**(10): 4007-4021.

Ong, S. E., B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey and M. Mann (2002). "Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics." Mol Cell Proteomics **1**(5): 376-386.

Ong, S. E. and M. Mann (2005). "Mass spectrometry-based proteomics turns quantitative." Nat Chem Biol **1**(5): 252-262.

Orchard, S., J. P. Albar, E. W. Deutsch, P. A. Binz, A. R. Jones, D. Creasy and H. Hermjakob (2008). “Annual spring meeting of the Proteomics Standards Initiative 23-25 April 2008, Toledo, Spain.” Proteomics **8**(20): 4168-4172.

Orchard, S., J. P. Albar, E. W. Deutsch, M. Eisenacher, P. A. Binz and H. Hermjakob (2010). “implementing data standards: a report on the HUPOPSI workshop September 2009, Toronto, Canada.” Proteomics **10**(10): 1895-1898.

Orchard, S., J. P. Albar, E. W. Deutsch, M. Eisenacher, J. A. Vizcaino and H. Hermjakob (2011). “Enabling BioSharing - a report on the Annual Spring Workshop of the HUPO-PSI April 11-13, 2011, EMBL-Heidelberg, Germany.” Proteomics **11**(22): 4284-4290.

Orchard, S., J. P. Albar, E. W. Deutsch, M. Eisenacher, P. A. Binz, S. Martinez-Bartolome, J. A. Vizcaino and H. Hermjakob (2012). “From proteomics data representation to public data flow: a report on the HUPO-PSI workshop September 2011, Geneva, Switzerland.” Proteomics **12**(3): 351-355.

Orchard, S., R. Apweiler, R. Barkovich, D. Field, J. S. Garavelli, D. Horn, A. Jones, P. Jones, R. Julian, R. McNally, J. Nerothin, N. Paton, A. Pizarro, S. Seymour, C. Taylor, S. Wiemann and H. Hermjakob (2006). “Proteomics and Beyond: a report on the 3rd Annual Spring Workshop of the HUPO-PSI 21-23 April 2006, San Francisco, CA, USA.” Proteomics **6**(16): 4439-4443.

Orchard, S., P. A. Binz and H. Hermjakob (2009). “Second Joint HUPO publication and Proteomics Standards Initiative workshop.” Proteomics **9**(19): 4426-4428.

Orchard, S., E. W. Deutsch, P. A. Binz, A. R. Jones, D. Creasy, L. Montechi-Palazzi, G. Corthals and H. Hermjakob (2009). “Annual spring meeting of the Proteomics Standards Initiative.” Proteomics **9**(19): 4429-4432.

Orchard, S., H. Hermjakob and R. Apweiler (2003). “The proteomics standards initiative.” Proteomics **3**(7): 1374-1376.

Orchard, S., H. Hermjakob, P. A. Binz, C. Hoogland, C. F. Taylor, W. Zhu, R. K. Julian, Jr. and R. Apweiler (2005). “Further steps towards data standardisation: the Proteomic Standards Initiative HUPO 3(rd) annual congress, Beijing 25-27(th) October, 2004.” Proteomics **5**(2): 337-339.

Orchard, S., H. Hermjakob, C. F. Taylor, F. Potthast, P. Jones, W. Zhu, R. K. Julian, Jr. and R. Apweiler (2005). “Further steps in standardisation. Report of the second annual Proteomics Standards Initiative Spring Workshop (Siena, Italy 17-20th April 2005).” Proteomics **5**(14): 3552-3555.

Orchard, S., H. Hermjakob, C. Taylor, P. A. Binz, C. Hoogland, R. Julian, J. S. Garavelli, R. Aebersold and R. Apweiler (2006). “Autumn 2005 Workshop of the Human Proteome Organisation Proteomics Standards Initiative (HUPO-PSI) Geneva, September, 4-6, 2005.” Proteomics **6**(3): 738-741.

Orchard, S., C. Hoogland, A. Bairoch, M. Eisenacher, H. J. Kraus and P. A. Binz (2009). “Managing the data explosion. A report on the HUPO-PSI Workshop. August 2008, Amsterdam, The Netherlands.” Proteomics **9**(3): 499-501.

Orchard, S., A. Jones, J. P. Albar, S. Y. Cho, K. H. Kwon, C. Lee and H. Hermjakob (2010). “Tackling quantitation: a report on the annual Spring Workshop of the HUPO-PSI 28-30 March 2010, Seoul, South Korea.” Proteomics **10**(17): 3062-3066.

Orchard, S., P. Kersey, H. Hermjakob and R. Apweiler (2003). “The HUPO Proteomics Standards Initiative Meeting: Towards Common Standards for Exchanging Proteomics Data.” Comp Funct Genomics **4**(1): 16-19.

Orchard, S., P. Kersey, W. Zhu, L. Montecchi-Palazzi, H. Hermjakob and R. Apweiler (2003). “Progress in Establishing Common Standards for Exchanging Proteomics Data: The Second Meeting of the HUPO Proteomics Standards Initiative.” Comp Funct Genomics **4**(2): 203-206.

Orchard, S., L. Martens, J. Tasman, P. A. Binz, J. P. Albar and H. Hermjakob (2008). “6th HUPO Annual World Congress - Proteomics Standards Initiative Workshop 6-10 October 2007, Seoul, Korea.” Proteomics **8**(7): 1331-1333.

Orchard, S., L. Montechi-Palazzi, E. W. Deutsch, P. A. Binz, A. R. Jones, N. Paton, A. Pizarro, D. M. Creasy, J. Wojcik and H. Hermjakob (2007). "Five years of progress in the Standardization of Proteomics Data 4th Annual Spring Workshop of the HUPO-Proteomics Standards Initiative April 23-25, 2007 Ecole Nationale Superieure (ENS), Lyon, France." Proteomics **7**(19): 3436-3440.

Orchard, S. and P. Ping (2009). "HUPO World Congress Publication Committee meeting. August 2008, Amsterdam, The Netherlands." Proteomics **9**(3): 502-503.

Orchard, S., L. Salwinski, S. Kerrien, L. Montecchi-Palazzi, M. Oesterheld, V. Stumpflen, A. Ceol, A. Chatr-aryamontri, J. Armstrong, P. Woollard, J. J. Salama, S. Moore, J. Wojcik, G. D. Bader, M. Vidal, M. E. Cusick, M. Gerstein, A. C. Gavin, G. Superti-Furga, J. Greenblatt, J. Bader, P. Uetz, M. Tyers, P. Legrain, S. Fields, N. Mulder, M. Gilson, M. Niepmann, L. Burgoon, J. De Las Rivas, C. Prieto, V. M. Perreau, C. Hogue, H. W. Mewes, R. Apweiler, I. Xenarios, D. Eisenberg, G. Cesareni and H. Hermjakob (2007). "The minimum information required for reporting a molecular interaction experiment (MIMIx)." Nat Biotechnol **25**(8): 894-898.

Orchard, S., C. F. Taylor, P. Jones, L. Montechi-Palazzo, P. A. Binz, A. R. Jones, A. Pizarro, R. K. Julian, Jr. and H. Hermjakob (2007). "Entering the implementation era: a report on the HUPO-PSI Fall workshop 25-27 September 2006, Washington DC, USA." Proteomics **7**(3): 337-339.

Orchard, S., W. Zhu, R. K. Julian, Jr., H. Hermjakob and R. Apweiler (2003). "Further advances in the development of a data interchange standard for proteomics data." Proteomics **3**(10): 2065-2066.

Pappin, D. J., P. Hojrup and A. J. Bleasby (1993). "Rapid identification of proteins by peptide-mass fingerprinting." Curr Biol **3**(6): 327-332.

Paradela, A., S. B. Bravo, M. Henriquez, G. Riquelme, F. Gavilanes, J. M. Gonzalez-Ros and J. P. Albar (2005). "Proteomic analysis of apical microvillous membranes of syncytiotrophoblast cells reveals a high degree of similarity with lipid rafts." J Proteome Res **4**(6): 2435-2441.

Paradela, A., P. R. Escuredo and J. P. Albar (2006). "Geographical focus. Proteomics initiatives in Spain: ProteoRed." Proteomics **6 Suppl 2**: 73-76.

Pedrioli, P. G., J. K. Eng, R. Hubley, M. Vogelzang, E. W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R. H. Angeletti, R. Apweiler, K. Cheung, C. E. Costello, H. Hermjakob, S. Huang, R. K. Julian, E. Kapp, M. E. McComb, S. G. Oliver, G. Omenn, N. W. Paton, R. Simpson, R. Smith, C. F. Taylor, W. Zhu and R. Aebersold (2004). "A common open representation of mass spectrometry data and its application to proteomics research." Nat Biotechnol **22**(11): 1459-1466.

Perkins, D. N., D. J. Pappin, D. M. Creasy and J. S. Cottrell (1999). "Probability-based protein identification by searching sequence databases using mass spectrometry data." Electrophoresis **20**(18): 3551-3567.

Picotti, P., B. Bodenmiller, L. N. Mueller, B. Domon and R. Aebersold (2009). "Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics." Cell **138**(4): 795-806.

Price, P. (1991). "Standard definitions of terms relating to mass spectrometry." Journal of the American Society for Mass Spectrometry **2**(4): 12.

Prieto, G., K. Aloria, N. Osinalde, A. Fullaondo, J. M. Arizmendi and R. Matthiesen (2012). "PAnalyzer: a software tool for protein inference in shotgun proteomics." BMC Bioinformatics **13**: 288.

Rabilloud, T. (2013). "When 2D is not enough, go for an extra dimension." Proteomics **13**(14): 2065-2068.

Rodriguez-Perez, M. A., A. Medina-Aunon, S. M. Encarnacion-Guevara, S. Bernal-Silvia, H. Barrera-Saldana and J. P. Albar-Ramirez (2008). "In silico analysis of protein neoplastic biomarkers for cervix and uterine cancer." Clin Transl Oncol **10**(10): 604-617.

Roepstorff, P. and J. Fohlman (1984). "Proposal for a common nomenclature for sequence ions in mass spectra of

peptides." Biomed Mass Spectrom **11**(11): 601.

Saiki, R., D. Gelfand, S. Stoffel, S. Scharf, R. Higuchi, G. Horn, K. Mullis and H. Erlich (1988). "Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase." Science **239**(4839): 487-491.

Schmidt, A., B. Bisle and T. Kislinger (2009). "Quantitative peptide and protein profiling by mass spectrometry." Methods Mol Biol **492**: 21-38.

Segura, V., J. A. Medina-Aunon, E. Guruceaga, S. I. Gharbi, C. Gonzalez-Tejedo, M. M. Sanchez del Pino, F. Canals, M. Fuentes, J. I. Casal, S. Martinez-Bartolome, F. Elortza, J. M. Mato, J. M. Arizmendi, J. Abian, E. Oliveira, C. Gil, F. Vivanco, F. Blanco, J. P. Albar and F. J. Corrales (2013). "Spanish human proteome project: dissection of chromosome 16." J Proteome Res **12**(1): 112-122.

Shevchenko, A., M. Wilm, O. Vorm and M. Mann (1996). "Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels." Anal Chem **68**(5): 850-858.

Stahl-Zeng, J., V. Lange, R. Ossola, K. Eckhardt, W. Krek, R. Aebersold and B. Domon (2007). "High sensitivity detection of plasma proteins by multiple reaction monitoring of N-glycosites." Mol Cell Proteomics **6**(10): 1809-1817.

States, D. J., G. S. Omenn, T. W. Blackwell, D. Fermin, J. Eng, D. W. Speicher and S. M. Hanash (2006). "Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study." Nat Biotechnol **24**(3): 333-338.

Syka, J. E., J. J. Coon, M. J. Schroeder, J. Shabanowitz and D. F. Hunt (2004). "Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry." Proc Natl Acad Sci U S A **101**(26): 9528-9533.

Taylor, C. F., P. A. Binz, R. Aebersold, M. Affolter, R. Barkovich, E. W. Deutsch, D. M. Horn, A. Huhmer, M. Kussmann, K. Lilley, M. Macht, M. Mann, D. Muller, T. A. Neubert, J. Nickson, S. D. Patterson, R. Raso, K. Resing, S. L. Seymour, A. Tsugita, I. Xenarios, R. Zeng and R. K. Julian, Jr. (2008). "Guidelines for reporting the use of mass spectrometry in proteomics." Nat Biotechnol **26**(8): 860-861.

Taylor, C. F., N. W. Paton, K. L. Garwood, P. D. Kirby, D. A. Stead, Z. Yin, E. W. Deutsch, L. Selway, J. Walker, I. Riba-Garcia, S. Mohammed, M. J. Deery, J. A. Howard, T. Dunkley, R. Aebersold, D. B. Kell, K. S. Lilley, P. Roepstorff, J. R. Yates, 3rd, A. Brass, A. J. Brown, P. Cash, S. J. Gaskell, S. J. Hubbard and S. G. Oliver (2003). "A systematic approach to modeling, capturing, and disseminating proteomics experimental data." Nat Biotechnol **21**(3): 247-254.

Taylor, C. F., N. W. Paton, K. S. Lilley, P. A. Binz, R. K. Julian, Jr., A. R. Jones, W. Zhu, R. Apweiler, R. Aebersold, E. W. Deutsch, M. J. Dunn, A. J. Heck, A. Leitner, M. Macht, M. Mann, L. Martens, T. A. Neubert, S. D. Patterson, P. Ping, S. L. Seymour, P. Souda, A. Tsugita, J. Vandekerckhove, T. M. Vondriska, J. P. Whitelegge, M. R. Wilkins, I. Xenarios, J. R. Yates, 3rd and H. Hermjakob (2007). "The minimum information about a proteomics experiment (MIAPE)." Nat Biotechnol **25**(8): 887-893.

Thompson, A., J. Schafer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, R. Johnstone, A. K. Mohammed and C. Hamon (2003). "Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS." Anal Chem **75**(8): 1895-1904.

Uhlen, M., E. Bjorling, C. Agaton, C. A. Szigyarto, B. Amini, E. Andersen, A. C. Andersson, P. Angelidou, A. Asplund, C. Asplund, L. Berglund, K. Bergstrom, H. Brumer, D. Cerjan, M. Ekstrom, A. Elobeid, C. Eriksson, L. Fagerberg, R. Falk, J. Fall, M. Forsberg, M. G. Bjorklund, K. Gumbel, A. Halimi, I. Hallin, C. Hamsten, M. Hansson, M. Hedhammar, G. Hercules, C. Kampf, K. Larsson, M. Lindskog, W. Lodewyckx, J. Lund, J. Lundeberg, K. Magnusson, E. Malm, P. Nilsson, J. Odling, P. Oksvold, I. Olsson, E. Oster, J. Ottosson, L. Paavilainen, A. Persson, R. Rimini, J. Rockberg, M. Runeson, A. Sivertsson, A. Skollermo, J. Steen, M. Stenvall, F. Sterky, S. Stromberg, M. Sundberg, H. Tegel, S. Tourle, E. Wahlund, A. Walden, J. Wan, H. Wernerus, J. Westberg, K. Wester, U. Wrethagen, L. L. Xu, S. Hober and F. Ponten (2005). "A human protein atlas for normal and cancer tissues based on antibody proteomics." Mol Cell Proteomics **4**(12): 1920-1932.

Uhlen, M., P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, S.

Hober, H. Wernerus, L. Bjorling and F. Ponten (2010). “Towards a knowledge-based Human Protein Atlas.” Nat Biotechnol **28**(12): 1248-1250.

UniProt_Consortium (2013). “Update on activities at the Universal Protein Resource (UniProt) in 2013.” Nucleic Acids Res **41**(Database issue): D43-47.

Unlu, M., M. E. Morgan and J. S. Minden (1997). “Difference gel electrophoresis: a single gel method for detecting changes in protein extracts.” Electrophoresis **18**(11): 2071-2077.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh and X. Zhu (2001). “The sequence of the human genome.” Science **291**(5507): 1304-1351.

Vizcaino, J. A., R. G. Cote, A. Csordas, J. A. Dianes, A. Fabregat, J. M. Foster, J. Griss, E. Alpi, M. Birim, J. Contell, G. O’Kelly, A. Schoenegger, D. Ovelleiro, Y. Perez-Riverol, F. Reisinger, D. Rios, R. Wang and H. Hermjakob (2013). “The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013.” Nucleic Acids Res **41**(Database issue): D1063-1069.

Walzer, M., D. Qi, G. Mayer, J. Uszkoreit, M. Eisenacher, T. Sachsenberg, F. F. Gonzalez-Galarza, J. Fan, C. Bessant, E. W. Deutsch, F. Reisinger, J. A. Vizcaino, J. A. Medina-Aunon, J. P. Albar, O. Kohlbacher and A. R. Jones (2013). “The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics.” Mol Cell Proteomics.

Wang, R., A. Fabregat, D. Rios, D. Ovelleiro, J. M. Foster, R. G. Cote, J. Griss, A. Csordas, Y. Perez-Riverol, F. Reisinger, H. Hermjakob, L. Martens and J. A. Vizcaino (2012). “PRIDE Inspector: a tool to visualize and validate MS proteomics data.” Nat Biotechnol **30**(2): 135-137.

Washburn, M. P., D. Wolters and J. R. Yates, 3rd (2001). “Large-scale analysis of the yeast proteome by multidimensional protein identification technology.” Nat Biotechnol **19**(3): 242-247.

Wasinger, V. C., S. J. Cordwell, A. Cerpa-Poljak, J. X. Yan, A. A. Gooley, M. R. Wilkins, M. W. Duncan, R. Harris, K. L. Williams and I. Humphery-Smith (1995). “Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium.” Electrophoresis **16**(7): 1090-1094.

Weston, A. D. and L. Hood (2004). “Systems biology, proteomics, and the future of health care: Toward predictive, preventative, and personalized medicine.” Journal of Proteome Research **3**(2): 179-196.

Wheeler, D. A., M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y. J. Chen, V. Makhijani, G. T. Roth, X. Gomes, K. Tartaro, F. Niazi, C. L. Turcotte, G. P. Irzyk, J. R. Lupski, C. Chinault, X. Z. Song, Y. Liu, Y. Yuan, L. Nazareth, X. Qin, D. M. Muzny, M. Margulies, G. M. Weinstock, R. A. Gibbs and J. M. Rothberg (2008). “The complete genome of an individual by massively parallel DNA sequencing.” Nature **452**(7189): 872-876.

Wiese, S., K. A. Reidegeld, H. E. Meyer and B. Warscheid (2007). “Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research.” Proteomics **7**(3): 340-350.

Wilkins, M. R., E. Gasteiger, C. H. Wheeler, I. Lindskog, J. C. Sanchez, A. Bairoch, R. D. Appel, M. J. Dunn and D. F. Hochstrasser (1998). “Multiple parameter cross-species protein identification using MultiIdent--a world-wide web accessible tool.” Electrophoresis **19**(18): 3199-3206.

Yan, J. X., R. A. Harry, C. Spibey and M. J. Dunn (2000). “Postelectrophoretic staining of proteins separated by two-dimensional gel electrophoresis using SYPRO dyes.” Electrophoresis **21**(17): 3657-3665.

Yates, J. R., C. I. Ruse and A. Nakorchevsky (2009). “Proteomics by mass spectrometry: approaches, advances, and applications.” Annu Rev Biomed Eng **11**: 49-79.

Zeeberg, B. R., W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett and J. N. Weinstein (2003). “GoMiner: a resource for biological interpretation of genomic and proteomic data.” Genome Biol **4**(4): R28.

Zhang, B., S. Kirov and J. Snoddy (2005). “WebGestalt: an integrated system for exploring gene sets in various biological contexts.” Nucleic Acids Res **33**(Web Server issue): W741-748.

Zhang, B., D. Schmoyer, S. Kirov and J. Snoddy (2004). “GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies.” BMC Bioinformatics **5**: 16.

Zhang, X., A. Fang, C. P. Riley, M. Wang, F. E. Regnier and C. Buck (2010). “Multi-dimensional liquid chromatography in proteomics--a review.” Anal Chim Acta **664**(2): 101-113.

Zubarev, R. A., D. M. Horn, E. K. Fridriksson, N. L. Kelleher, N. A. Kruger, M. A. Lewis, B. K. Carpenter and F. W. McLafferty (2000). “Electron capture dissociation for structural characterization of multiply charged protein cations.” Anal Chem **72**(3): 563-573.

Anexo 1. Tabla de aminoácidos y sus propiedades

Aminoácido	Abrev	Símb	Fórmula Molecular	Cadena lateral	MW	PI	Clasificación	Esen**
Glicina	Gly	G	$C_2H_5NO_2$	Hidrófilo	75.07	6.06	Alifático	No
Alanina	Ala	A	$C_3H_7NO_2$	Hidrófobo	89.09	6.11	Alifático	No
Valina	Val	V	$C_5H_{11}NO_2$	Hidrófobo	117.15	6.00	Alifático	Si
Leucina	Leu	L	$C_6H_{13}NO_2$	Hidrófobo	131.17	6.01	Alifático	Si
Isoleucina	ILe	I	$C_6H_{13}NO_2$	Hidrófobo	131.17	6.05	Alifático	Si
Fenilalanina	Phe	F	$C_9H_{11}NO_2$	Hidrófobo	165.19	5.49	Aromático	Si
Tirosina	Tyr	Y	$C_9H_{11}N_1O_3$	Hidrófilo	181.19	5.64	Aromático	No
Triptófano	Trp	W	$C_{11}H_{12}N_2O_2$	Hidrófobo	204.23	5.89	Aromático	Si
Serina	Ser	S	$C_3H_7NO_3$	Hidrófilo	105.09	5.68	Alcohol	No
Treonina	Thr	T	$C_4H_9NO_3$	Hidrófilo	119.12	5.60	Alcohol	Si
Cisteína	Cys	C	$C_3H_7NO_2S$	Hidrófilo	121.16	5.05	Con Azufre	No
Metionina	Met	M	$C_5H_{11}NO_2S$	Hidrófobo	149.21	5.74	Con Azufre	Si
A. AspárticO	Asp	D	$C_4H_7NO_4$	Ácido	133.10	2.85	Ácido	No
A. Glutámico	Glu	E	$C_5H_9NO_4$	Ácido	147.13	3.15	Ácido	No
Histidina	His	H	$C_6H_9N_3O_2$	Básico	155.16	7.60	Básico	Si

Lisina	Lys	K	$C_6H_{14}N_2O_2$	Básico	146.19	9.60	Básico	Si
Arginina	Arg	R	$C_6H_{14}N_4O_2$	Básico	174.20	10.76	Básico	Si
Asparagina	Asn	N	$C_4H_8N_2O_3$	Hidrófilo	132.12	5.41	Amida	No
Glutamina	Gln	Q	$C_5H_{10}N_2O_3$	Hidrófilo	146.15	5.65	Amida	No
Prolina	Pro	P	$C_5H_9NO_2$	Hidrófobo	115.13	6.30	Inminoácido*	No

* También es alifático.

** Se llaman esenciales porque el organismo en cuestión no puede sintetizarlos de manera propia y por tanto hay que suministrarlos en la dieta.

BLOQUE 3:
COPIA PUBLICACIONES